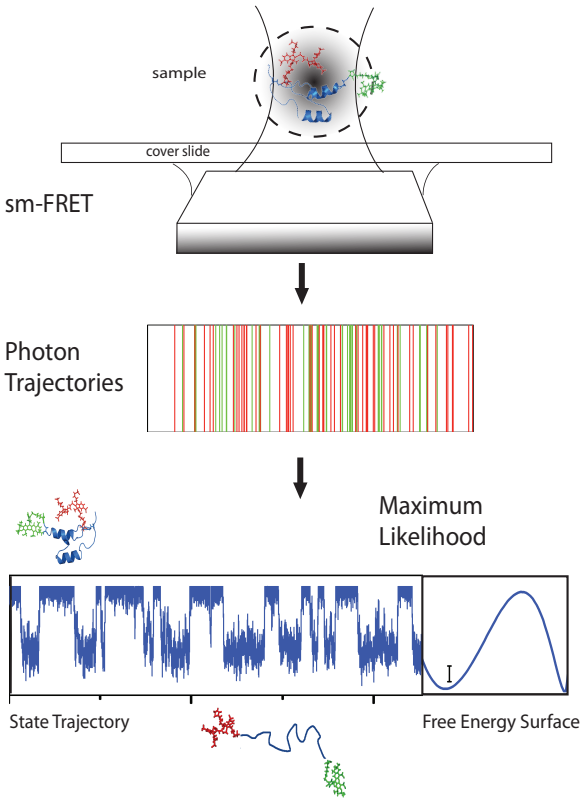
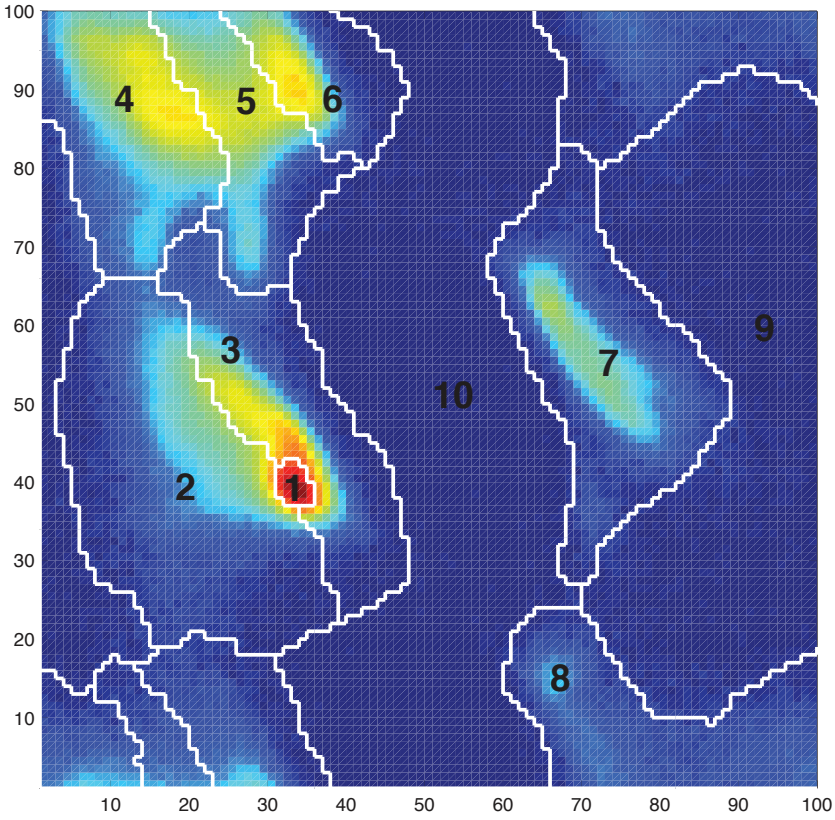
Universidad Autonoma de Madrid

Facultidad de Ciencias

Programa de Doctorado en Biofisica

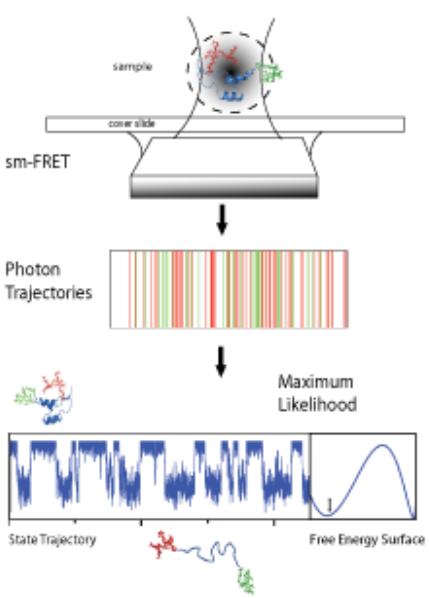


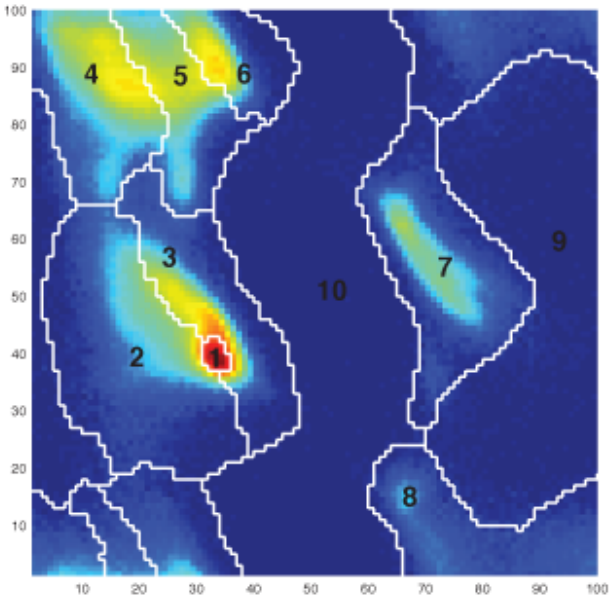


Stochastic and Statistical Analysis For Investigating Protein Folding Kinetics

Ravishankar Ramanathan

Universidad Autonoma de Madrid
Facultidad de Ciencias
Programa de Doctorado en Biofisica





Stochastic and Statistical Analysis For Investigating Protein Folding Kinetics

Ravishankar Ramanathan

Stochastic and Statistical Analyses for Investigating Protein Folding Kinetics

A Thesis Submitted to the
Faculty of Sciences of
Universidad Autonoma de Madrid

By

Ravishankar Ramanathan

in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

December 2014

Stochastic and Statistical Analysis for Investigating Protein Folding Kinetics

By
Ravishankar Ramanathan

Thesis Director:
Prof. Victor Muñoz

Dedication

To all those who strive to appreciate and love life always just for what it is,
not what it was or not what it will be

Acknowledgement

I, foremost want to acknowledge and thank my supervisor Victor Muñoz. Besides the brilliant mind he is in science, it was delightful for me to have various expansive chats with him over a wide range of topics, from economics to theism to history. He has supported me over all these years and even took the efforts of taking courses on financing and accounting for the entrepreneurship project. I'm very happy our lives have had this overlap for the past few years and that I had the chance to work with him.

The efficient cause of my coming to Victor's group in Spain was my friend, Athi whom I acknowledge here specifically for being so. We have known each other for over 17 years now, since the day I met him on the very first counseling for our undergrad studies. Besides the many memories over our long friendship, Madrid has made a life altering impact on both of us! But for him, my Madrid stay simply wouldn't have happened.

I express my gratitude for each and everyone who wished me well. I specifically thank Mourad who helped me a lot in the beginning with the settling in and with the tortuous Spanish bureaucracy when everything was new to me at Madrid. His helpful nature will always be appreciated. I thank David who sharing the garage like office that was allocated for us in the beginning that we called as the 'Cluster room' and also helping me in finding my first apartment to rent in Madrid was very supportive. Abhinav, another 'Cluster' company in the true sense – both in the office and in the computing clusters – Agamemnon and Agaricus, has specific thanks from me for all the conversations we had, interesting tech feeling we shared and the collaborations on our projects. I thank my other main collaborator Luis who besides all the good FRET data he provided, has also shared nice tennis sessions and trips together.

I want to thank my friend Malwina for all her help, support and motivation. And buddy Satish for his appreciation and support.

I acknowledge the support of and good times with all my other colleagues and Victor lab members, past and present. Sr. Jörg, Celia, Aga, Mila, Michele, Lorenzo, Nacho, Tanay, Raj, Clara, Martha and Eva, all have given me memorable times in my Spanish stay.

I also want to acknowledge the funding sources for my project – Marie Curie Grant and Prodestech Grant that supported and funded the projects I have participated in.

I deeply thank my family for all the love and support they have provided over these many years. Though separated by distance their unaltered, unalloyed affection has always boosted and bolstered my feelings and emotions. My dad and mom are really special and exceptional people and I'm blessed to be a part of my family. All I'm and will ever be is owed to them.

Abstract

Understanding how proteins are able to perform the multiple roles and activities they normally do is very important for our understanding of life. How or Why proteins adopt particular conformational states that facilitate their functionality is still an open-ended question. We have made enormous strides of progress towards figuring out the physico-chemical basis of this process and in general with our understanding of proteins. Continuous efforts in experimental, computational and theoretical approaches have enabled us to decipher the properties and behavior of this class of biomolecules considered to be the hardest nut to crack in the puzzle that is life. Now, we are in a firm footing with a solid theoretical framework in the form of the Energy Landscape Theory that offers the foundation on which to build scaffolds for the excavation of the mysteries of proteins. Computer Simulations have reached sufficient speeds and resolutions enabling us to come at this problem from a totally different side. Experimental approaches to study protein folding has arrived to the arenas of capturing single molecules in action as well as characterizing the crucial processes with ultrafast time resolution techniques. Convergence of these different approaches is at the forefront now with efforts toward iterative verification of computational results with experiments and replication of experimental results with simulations and a resultant net mutual learning. Towards this convergence, new methods and approaches of analysis are being developed that enables quantitative understanding of the data be it experimental or simulated, offering and incorporating simple yet fundamental views of the underlying physical processes. In this thesis, I present two such efforts that connect theory, simulations and experimental results.

Proteins being inherently subjected to stochastic forces and motions, I combine stochastic kinetic simulations with very simple models to elucidate and unravel their behavior and dynamics as tuned by their energetics and kinetic barriers. How the presence or absence of a barrier (even $\sim$ 1RT) marks a fundamental difference in the properties of proteins are clearly elucidated by analyzing the stochastic trajectories of single molecules. Firstly, I apply these simulations to study elementary helix-coil kinetics followed by the studies of barrier effects on protein folding. Simple stochastic kinetic simulations open a window to peer into the dynamics and behaviors of protein molecules and serves as a bridge between simple theoretical models and experiments and simulations. Later, I build a rigorous procedure based on maximum likelihood analysis to extract conformation dynamics from single molecule experiments on proteins. The method offers a quantitative way to analyze the measurements from time-resolved single molecule FRET experiments that are a leading tool in our arsenal to understand protein folding. By enabling to distinguish protein thermodynamics as well as simultaneously characterizing the dynamics of the underlying process, the method offers a robust and powerful approach to interpret time-stamped photon trajectory data and identify the right protein folding scenario that results in such data.

The second effort is a statistical approach to making connections between thermodynamics and protein structure. By utilizing the treasure trove of structural data from numerous X-ray crystallographic and NMR experiments available in the Protein Data Bank, we develop a method for extracting entropic costs of protein folding. We first develop a novel clustering methodology for partitioning the torsional angle space of protein backbones that is based on the statistics of backbone dihedral angles and reflects the natural preferences of individual amino acids to populate these particular regions. We introduce the side chain contributions based on rotameric distributions. Using a simple approach based on statistical thermodynamics, we then calculate the entropy cost of protein folding while calibrating and benchmarking it extensively with experimental data. We obtain a high correlation (R = 0.98) for the predicted and experimentally measured total entropic costs of folding. Comparisons of per residue entropy costs obtained after eliminating the well-known size scaling effects in protein folding establishes the high level of signal in our predictions. Using this approach, we make connections between a protein structure and its thermodynamics of folding. The structure based protein entropies are then introduced into a model of protein folding to improve its predictive capabilities.

These efforts combined, advance the recent attempts to build a convergence in the application of computational and experimental methods in expanding our understanding of protein folding.

Resumen

La comprensión de cómo las proteínas son capaces de abarcar los múltiples roles y actividades que desarrollan es muy importante para conocer el funcionamiento a nivel molecular de la vida. El cómo o el por qué las proteínas adoptan los estados conformacionales específicos que permiten su funcionalidad es una cuestión aún abierta. Se han logrado enormes avances para la resolución de las bases físico-químicas del proceso y hacia el entendimiento general de las proteínas. Los constantes esfuerzos experimentales, computacionales y teóricos han permitido descifrar las propiedades y el comportamiento de este tipo de biomoléculas, consideradas como la pieza más complicada de resolver en el puzle de la vida. Hoy en día, se han conseguido establecer bases sólidas en el marco teórico a través de la Teoría de los Paisajes Energéticos que ofrece un punto de partida sobre la cual construir andamiajes para alcanzar el conocimiento de los misterios de las proteínas. Las Simulaciones Computacionales han conseguido suficiente velocidad y resolución para permitirnos abordar el tema desde un punto de vista totalmente diferente. Los abordajes experimentales para estudiar el plegamiento de proteínas han logrado avanzar hasta alcanzar el seguimiento de moléculas únicas en acción, así como caracterizar procesos cruciales mediante técnicas con tiempos de resolución ultrarrápidos. Actualmente, la convergencia de estos diferentes abordajes constituye la vanguardia de este área investigadora, con esfuerzos dirigidos hacia la verificación iterativa de resultados computacionales con experimentos y replicación de los resultados experimentales con simulaciones, con la consiguiente red mutua de aprendizaje. Hacia esta convergencia están siendo enfocados los nuevos métodos y abordajes de análisis en desarrollo. Estos permiten la comprensión cuantitativa de los datos experimentales o simulados, ofreciendo e incorporando visiones fundamentales de los procesos físicos subyacentes. En esta tesis, presentaré dos de tales esfuerzos que conectan la teoría, las simulaciones y los resultados experimentales.

Estando las proteínas sometidas de forma inherente a fuerzas y movimientos estocásticos, he combinado simulaciones de cinética estocástica con modelos muy simples para elucidar y resolver, mediante el análisis de su energética y de sus barreras energéticas, el comportamiento y la dinámica que presentan. La presencia o ausencia de una barrera energética (incluso del orden de 1 kT) marca una diferencia fundamental en las propiedades de las proteínas, hecho que es claramente elucidado mediante el análisis de trayectorias estocásticas de moléculas únicas. Primero, he aplicado estas simulaciones al estudio cinético de la transición elemental hélice-ovillo, seguido por la aplicación al estudio del efecto de cambios de la barrera energética en el plegamiento de proteínas. Las simulaciones de cinética estocástica simples abren la posibilidad de mirar de cerca la dinámica y el comportamiento de moléculas proteicas y sirven de puente entre modelos teóricos simples y datos procedentes de experimentos o de simulaciones. Posteriormente, he creado un procedimiento riguroso basado en un análisis de máxima probabilidad para extraer información de la dinámica conformacional a partir de experimentos de molécula única de proteínas. El

método ofrece un medio cuantitativo de analizar las medidas de experimentos de FRET de molécula única, técnica que se ha convertido en una herramienta puntera en nuestro arsenal para entender el plegamiento de las proteínas. Gracias a la posibilidad de caracterizar la termodinámica de las proteínas así como la dinámica del proceso subyacente, el método ofrece una aproximación robusta y poderosa para interpretar los datos de trayectorias de fotones con precisión temporal generadas por una molécula proteica única e identificar el correcto escenario de plegamiento proteico que produce esos datos.

El segundo esfuerzo engloba la realización de una aproximación estadística para hacer conexiones entre la termodinámica y la estructura de una proteína. Mediante el uso de la inapreciable colección de datos estructurales procedentes de numerosos experimentos de cristalografía de rayos X y de RMN disponibles en el banco de datos de proteínas (PDB), hemos desarrollado un método para extraer el coste entrópico que supone el plegamiento de una proteína. En un primer paso, hemos desarrollado una nueva metodología de agrupamiento para dividir el rango de valores de ángulos de torsión de la cadena principal que está basada en estadísticas de los ángulos diedros de la cadena principal de proteínas con estructura conocida y que refleja las preferencias naturales de aminoácidos individuales para ocupar dichas divisiones. Hemos añadido la contribución de las cadenas laterales de los aminoácidos basándonos en la distribución de rotámeros. Mediante el uso de aproximaciones simples, basadas en termodinámica estadística, hemos calculado el coste entrópico del plegamiento de proteínas, para posteriormente calibrar y evaluar estos valores con datos experimentales. Hemos obtenido una correlación alta ($R = 0.98$) para los costes entrópicos totales del plegamiento predichos y medidos experimentalmente. La comparación con datos previamente publicados del coste entrópico por residuo obtenido tras eliminar los efectos bien conocidos de escalado por tamaño en el plegamiento de proteínas establece el alto nivel de señal en nuestras predicciones. Utilizando esta aproximación, hemos realizado conexiones entre la estructura de una proteína y su termodinámica de plegamiento. La entropía de una proteína basada en su estructura ha sido posteriormente introducida en un modelo de plegamiento para mejorar su capacidad de predicción.

Estos esfuerzos combinados suponen un avance dentro de los recientes intentos de construir una convergencia entre métodos computacionales y experimentales para expandir nuestro conocimiento sobre el plegamiento de proteínas.

Table Of Contents

List Of Figures

Abbreviations

DSC	Differental scanning calorimetry
FRET	Förster resonance energy transfer
smFRET	Single-molecule Förster resonance energy transfer
FES	Free Energy Surface
FEH	FRET efficiency histograms
ML	Maximum likelihood
MLA	Maximum likelihood analysis
T-jump	Temperature jump
R	Universal Gas Constant

1 Introduction

Life is an exquisite process and as we know it, is made possible by a great level of self-assembly of and coordination between different macromolecules and molecular components. It is marked with active and nonstop energetic efforts against the entropic forces of the universe that are constantly pushing any living system towards the equilibrium state of its simple constituent matter. This equilibrium is otherwise called 'death' and life is a constant process, activity and striving for non-equilibrium and dynamics. Of the different criteria that have been commonly accepted to define 'Life', Self-Organization and Self-Replication forms the essence[1]. Self-organization is highly anti-entropic and implies a mechanism for processing of energy that has to come from and exchanged with the environment. Self-replication on the other hand, besides the thermodynamic implications, also necessitates a representative factor or a blueprint that is to be replicated and passed on. An effective coupling between these essential principles also necessitates a self-corrective mechanism and Darwinian evolution plays this role.

Deoxyribonucleic acid, DNA serves as the replicative blueprint for Life, as we know it. The information content for different living organisms is stored in their DNA that gets replicated and passed on. It has been unequivocally established that this information ultimately encodes, via an intermediary, the most important group of biomolecules – the Proteins. The intermediary biomolecules are RNA and they are also very important in their own right, but the functional genetic information in the DNA primarily codes for proteins.

Proteins are the centerpiece of dynamics and action in biological systems as they perform various functions from structural scaffolding of the cells, environmental response, energy processing and chemical inter-conversions to replication of the genetic material encoding the life forms.

Proteins are the pinnacles for self-assembly that Life has produced and have continued to marvel innumerable scientists across disciplines over many decades, with their feat of reversibly forming specific three-dimensional patterns and structures of their constituent atoms. In the 1800s, proteins were identified as the most important category of biomolecules by the seminal work of Johannes Mulder who first used the name 'Proteins'[2] that had been coined by Joseph Berzelius, a founder of modern chemistry based on the Greek word *proteos – the first* or *primary one.* They were further characterized in the early 1900s to be consisting of subunits called as amino acids, 20 of which have been identified to be standard and universal to all life. Proteins are biopolymers that are linear strings of *amino acids*, their monomeric units.

The fundamental and most important discovery in molecular biology by Watson and Crick[3] – DNA as the basis of heredity and further unraveling of the genetic code - the mechanism of how DNA acts as the blueprint of living forms has led to the conclusion that the information encoded is that for the proteins. In fact, the

central dogma of molecular biology *DNA->RNA->Proteins->DNA* could be seen in the light that DNA are carriers of information encoding the central players which are *proteins*. Mapping of 1-dimensional coding in the genetic code to the 1-dimensional strings of amino acids is very well established and indeed forms the basis of biotechnology. At the level of molecular biology, it is a solved problem. But the more interesting aspect of how the 1-dimensional strings of amino acids translated from the RNA by the ribosome machinery encodes the information for a specific 3-dimesional spatial pattern and structure of atoms, also referred to as the 'protein folding' problem is still not totally resolved. Over 50 years of research has resulted in a good handle and understanding of the fundamentals of this problem, with rapid strides having been made in the last 20 years. Nevertheless, we are still far from a complete picture of all the principles of protein folding.

In 1960s, Christian Anfinsen made a basic discovery about these exemplary self-assemblers through his work on an enzyme ribonuclease called as RNAse I, that proteins *in vitro* could be reversibly 'natured' after being denatured through agents such as temperature without the need for any other external factors or cellular components[4]. This discovery is one of the seminal ones in thwarting the 'vital force' hypothesis, laying another of the foundational stones for a physical and molecular underpinning to biology. Anfinsen also made the crucial conclusion called as 'thermodynamic postulate' according to which, under physiological conditions the natural state of such proteins must be the state with minimum free energy. Proteins in solution could thus be treated as any other chemical for further studies and experimentations and the cellular components that were assumed to play a role in the final assembly of proteins in fact do not and can be totally dispensed with. Many proteins have since been totally characterized *in vitro*, independent of their cellular environments and other factors they could encounter when *in vivo*. The behaviors and properties of proteins when *in vivo* have been predicted and in multiple cases confirmed to be similar to them being *in vitro,* the physical effects of macromolecular crowding in the very dense cellular environments being appropriately factored in. [5]

Proteins *in vivo* typically do not work in isolation but interact with many other factors such as other proteins, lipids, nucleotides or other small molecules to perform and deliver their cellular roles. But of themselves they just need to adopt the appropriate 3-dimensional structure to perform their functions. This decoupling between the synthesis machinery i.e. the route and mechanism of protein synthesis and its folding to the final structure and performance of its function is what has made possible the development of biotechnology. The information necessary for the final, functional state of small proteins is completely contained within the sequence of their monomeric units – the 1-dimensional protein sequence. The constituent amino acids could be stringed together through whichever synthetic route, either completely chemically or made biologically using the ribosomal machinery. Same organism where the protein originally comes from could be the biological source or any other organism typically *E.coli*, whose protein synthetic machinery could be commandeered into the synthesis of foreign gene sequences could produce the protein. In either case, the resulting protein will self-assemble into its final,

stable 3-D structure maintaining its functional activity. How a protein is able to do it is not yet totally understood and constitutes the main aspect of the protein folding problem.

1.1 3-D structure – molecular to mesoscopic systems

1.1.1 Descriptive vs Mechanistic Information

With many advances in gene sequencing technologies, the amount of genetic information of living organisms in the form of full genomic databases has grown exponentially in the past decade since the unraveling of human genome in 2001[6]. Nowadays, many other massive genome projects are underway and we have reached scales of 1000$ personal human genome sequencing that is already leading to routine genome sequencing of a large number of human individuals[7]. Genomic sequencing of many other organisms - domesticated animals such as cows, vital crops such as rice, majority of infectious microbes like *Yersinia pestis* of the bubonic plague, multiple viruses like the human immuno-deficiency virus (HIV) and the flu virus etc. has been completed and the genomic information is available. Now, we have graduated from genomics to meta-genomics, massively sequencing the microbiota in the human gut, microbes of the soil, seawater from different parts of the globe and any other environs teeming with bacterial life[8]. This explosive surge in gene sequencing has left us with an unprecedented amount of genetic information and data to deal with the first leg of the central dogma of molecular biology – the DNA.

With methods for profiling the RNA content of the cells using techniques such as RNAseq and microarrays, the dynamics of the genetic information at the second level of the central dogma is also being tackled quite effectively. Massive projects such as ENCODE[9] are transforming our understanding about this important level in biology. The discoveries of ribozymes, microRNAs, siRNAs have garnered RNA key importance over the past decade.

At the protein level, there has also been tremendous progress with protein sequencing, identification of modifications at particular sites of proteins *en masse*, protein-protein interaction networks, relative and absolute quantification of proteins at various locations inside the cells and structural details of proteins. Recently, *human proteome* has been mapped with the information of different types and abundance of proteins expressed in different cell types in the human body [10,11]. Given the direct mapping between DNA sequences and the encoded proteins, growth in DNA sequences has led to the concomitant surge in availability of translated protein sequence information. Different *Proteome* level data is now available for many organisms. Bioinformatics approaches for analysis of sequences and other information are well developed. These are used to understand as well as make predictions about structure, functions and dynamics from such data.

All such *descriptive* and to an extent dynamic information has led to the flourishing of fields such as '*systems biology*' that is fundamentally different from the traditional reductionist approaches adopted in biological scientific enquiry.

By integrating the static as well as dynamic information about the biological systems, quantitative predictions about their response to perturbations, about their future behavior based on initial conditions etc. are now possible. In the nascent field known as '*synthetic biology*' all such known descriptive and dynamic information are being integrated to catalog a list of standard biological parts that will serve to engineer different systems as routinely done in other manmade engineering disciplines such as mechanical and electrical engineering. Currently, with the developments in genome editing and DNA assembly, made to order living systems *a la carte* engineered for particular functions of interest are within the reach of biotechnology [12].

Nevertheless, in order to get into a totally quantitative biology, we still miss a big aspect of the puzzle of molecular biology viz., the protein folding problem.

1.1.2 Structure – function relationship

"If you want to understand function, study structure." - Francis Crick, Nobel Prize in Medicine, 1962.

"…Everything that living things do can be understood in terms of the jiggling and wiggling of atoms." – Richard Feynman, Nobel Prize in Physics, 1965

Structural biology – the mapping of atomic level structure of biomolecules has been fundamentally responsible for deciphering many molecular aspects of life. From the central discovery of the double helix of DNA to the structures of complex macromolecular machinery such as ribosomes it has ushered in atomic level description to our view and understanding of life.

The first atomic structure of a protein was that of myoglobin by Max Perutz in 1958, which opened up the field of protein structural biology. Since then, over hundred thousand protein structures have been experimentally determined using the techniques of X-ray crystallography and Nuclear Magnetic Resonance (NMR). Protein Data Bank (PDB)**,** the database of protein structures crossed the landmark *100,000* structures recently in May, 2014. Structure-function paradigm according to which the function of biomolecules could be understood and explained based on the 3-D spatial arrangements and patterns of their constituent atoms has been fundamental over the past decades evident from it garnering more than 14 Nobel Prizes from 1956 to 2014. Almost one in four Nobel Prizes in chemistry since 1956 has been awarded for biomolecular structure-related work. Structures of individual proteins and macromolecular complexes have shone light and offered understanding of their functioning. Yet again, the mechanisms of the formation of such orchestrated structural complexes and self-organized structures are still open problems pointing to the underlying grand 'protein folding problem'. The emergence of intrinsically disordered proteins [13] is another pointer to the gap in our current understanding of these dynamic biomolecules and how they render their various functions.

1.2 The Protein Folding Problem

The aspects of Protein folding problem are three fold:

1) Given a sequence of amino acids, what are the determinants of its 3-D structure, the so called folding code.
2) Even if one knows the final structure of a protein, what is the mechanism by which it reaches the structure?
3) Can we predict structure from just given a sequence? Given a structure (required for a particular function) how does one map out the sequence space that will result in that particular structure.

Enormous progress has been made in addressing all these aspects of the protein folding problem over the past 50 years, especially in the last two decades with the development of a consolidated framework of protein folding. Though the journal *Science*, in 2005 named the problem to be one among the 125 biggest unsolved problems in science[14], a notion has been mooted that broader and essential principles of the problem has already been cracked [15]. Such provocative claims were indeed made not to undermine the efforts to address the problem but to establish the point that what started as a specific research question has now spawned an entire field of protein physical science with multidisciplinary effort towards numerous research directions.

1.3 Energy Landscape Theory – the Framework

The reason protein folding has been regarded as very challenging is the issue of an astronomical number of conformations a protein might have to search in order to reach a specific native state. The mountain of experimental work on protein folding since 1950s has yielded a bewildering complexity of folding thermodynamics and kinetics results from a multiple variety of probes. In 1968, Cyrus Levinthal had proposed it as a paradox how a protein molecule is able to find its native state ever. Assuming 100 residues and 3 conformations per residue, resulting in $3^{100} = 5 \times 10^{47}$ conformations to search from, the protein molecule still ends up in the native conformation within tractable time. Even if the molecule takes 1 ps per conformation for the random search, it will take 10^{27} years for a single protein to fold, much more than the age of the universe. Yet, in reality proteins routinely fold in the order of seconds. The idea was that proteins have specific pathways of folding that they take en route to the native state and there was a general hunt for identifying and characterizing such pathways. The efforts to ferret out such pathways in all the diversity and details didn't succeed in identifying any. In general these efforts and the complexity of the behaviors identified from many experiments suffered from the lack of a solid theoretical framework. An alternative view that was statistical in nature and was based on the concepts from condensed matter physics and statistical thermodynamics[16] began to emerge in the late 1980s and mid 1990s from the groups of Peter Wolynes. According to the view, protein folding process is best understood by a statistical description of the protein's energy surface or the landscape and that folding occurs through organizing an ensemble of structures rather than through specifically defined structural intermediates. The main idea emerging from such

a view is that globally the protein folding energy landscape resembles a funnel but is somewhat rugged in shape, riddled with tiny traps where the molecule could reside transiently on its way to the bottom of the funnel, the native state that is an ensemble of conformations with low free energy. The energy landscape of the protein-solvent system is in an hyperdimensional space of all the atoms of the system. It is usually projected down and represented as a 3D surface with free energy as the vertical axis and the conformational degrees of freedom (entropy) as the horizontal axis.

Top of the funnel is populated with the completely unfolded conformations of the molecule with high entropies that essentially determines the width of the funnel. Height of the funnel is given by the net free energy change in the protein to reach its native state averaged over all the atoms including the solvent atoms. Funneled energy landscape of the protein facilitates folding through multiple microscopic routes rather than specific pathways. The unfolded molecule facilitated by stochastic thermal kicks could just 'flow or go' down the funnel taking any of the innumerable routes possible to reach the bottom low energy, low entropy native state with the loss of entropy compensated by the gain in energy. This in a way solves Levinthal's paradox by making it a non-paradox in essence. Energy Landscape view offers the most complete and quantitative picture of protein conformational space, including the fully folded native state, ensembles of various conformational substates near or far away from the native state, ensembles of folding intermediates (such as molten globules, collapsed states, transition states etc.) and the variously denatured and unfolded states. The folding funnel view is statistically encompassing all of the structural and energetic space of the protein molecule.

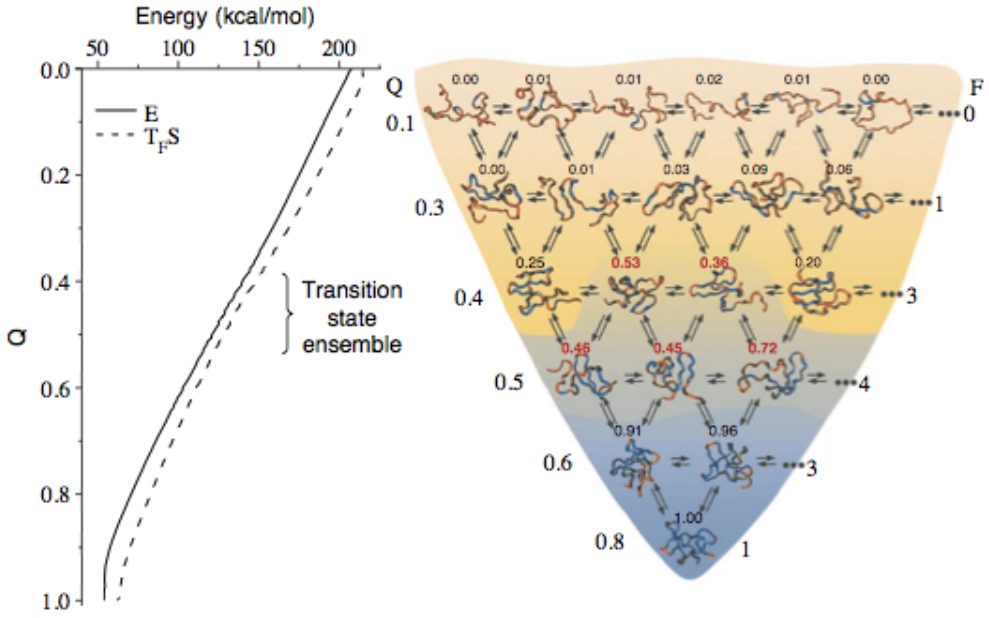


Figure 1.1 Sample Energy Lanscape of a protein. On the left, average energy and entropy of conformations with a particular value of the reaction coordinate Q (the fraction native contacts). On the right is a funnel diagram. Width of the funnel represents the entropy while the depth represented energy. The numbers on the sample conformations correspond to the probability to complete folding prior to unfolding. Entropy is favored at the top of the funnel and energy at its bottom. Figure adapted from page 57, Muñoz [17]

Energy landscape theory makes extensive comparisons between evolved natural proteins and random heteropolymers. It invokes the central '*principle of minimal frustration*' that states that naturally evolved proteins are optimized sequences of amino acids that have evolved to fold rapidly and efficiently to well-defined native states without being stuck at the many possible kinetic traps en route from the unfolded states. On the other hand, random sequences of amino acids possess large heterogeneity even in their native states with a very flat bottom of the funnel or possess a very rugged landscape with many local energy minima acting as kinetic traps. In such cases the random polymer is considered to be frustrated and their energy landscape rugged with their dynamics being referred to as 'glassy'.

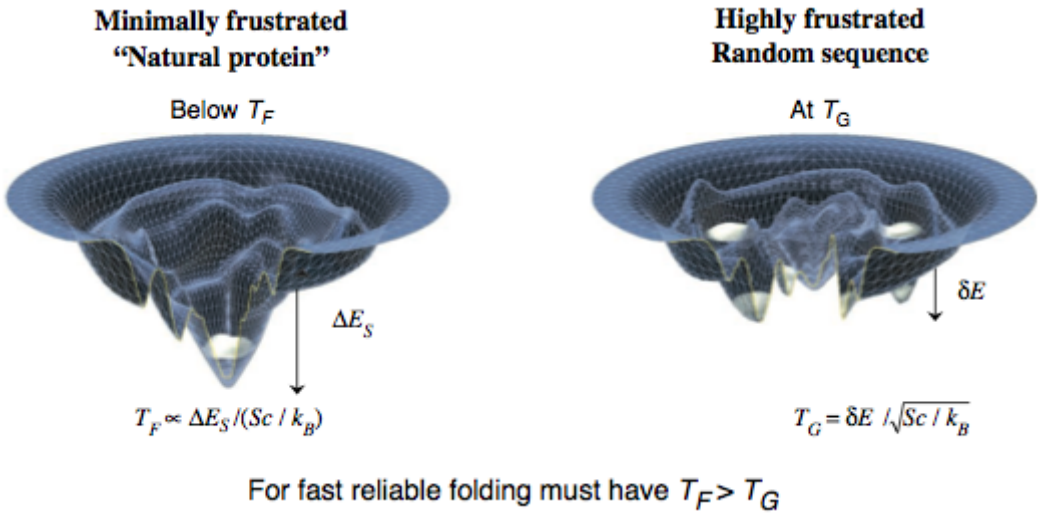


Figure 1.2 The left panel shows a minimally frustrated protein that having a clear low energy funnel bottom whereas the right panel is that of a highly frustrated random sequence. T_F is folding temperature and it depends on ΔE the energy gap between funnel minima and random states, and configurational entropy Sc. T_G is glass transition temperature that depends on root mean square fluctuation of collapsed random structural energy δE and configurational entropy at which few misfolded states dominate and act as traps. TF must be greater than TG for proteins to reliable fold. Figure adapted from page 59, Muñoz[17]

As the structural heterogeneity in the bottom of the funnel is low and the routes taken by the protein polymer to reach the native states are not strict, the funneled energy landscapes lends an explanation for the evolutionary and mutational robustness of natural proteins. Any single mutation may not alter the structure enough to take it completely out of the bottom of the funnel as potentially competing low energy states are still similar in structures. The funnel shaped energy landscapes of proteins thus lead to mutational robustness as well to environmental perturbations.

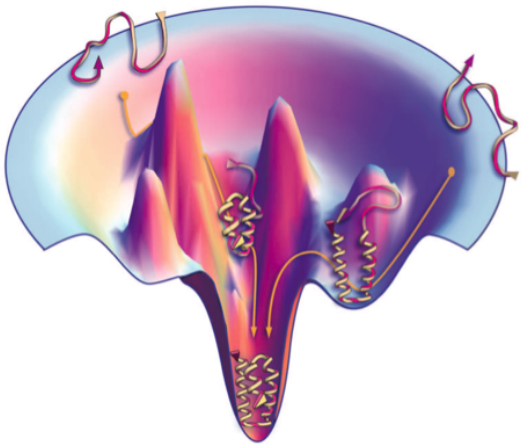


Figure 1.3 Typical schematic folding funnel of a small protein**.** Funnel with multiple high energy unfolded structures and single well-defined native minima. Folding occurs via multiple microscopic routes. Figure adapted from Dill, 2012[15]

1.4 Implications of Energy Landscape Theory

1.4.1 Low dimensional Projections

One of main results of the Energy Landscape theory is that the intractability of the hyperdimensional folding energy landscape of proteins could be solved by using low dimensional projections of the free energy using appropriately chosen tractable "reaction coordinates". Kinetics of the protein folding could be described as diffusion on this low-dimensional free energy surface a direct result from the Landscape theory. Onuchic and coworkers first verified the ability of a single reaction coordinate to capture all the essential details of folding in the analysis using computer simulations[18]. They studied the folding kinetics of a simplified model of protein, a 27-mer on a cubic lattice that itself is a complex system with $\sim 10^{17}$ possible conformations. Using a 1-dimensional projection of the free energy from the simulation against Q, the fraction of native contacts, they discovered that the mean folding time in the simulation could be calculated well just using Kramer's theory for a one-dimensional barrier crossing event. The calculation had no adjustable parameters and used information that was directly obtained from the simulation-barrier height from the free energy vs Q profile and the diffusion coefficient estimated using decay time of the autocorrelation function of the free energy profile. This rather surprising but key result shifted the emphasis to developing simple analytical models with free energy functions.

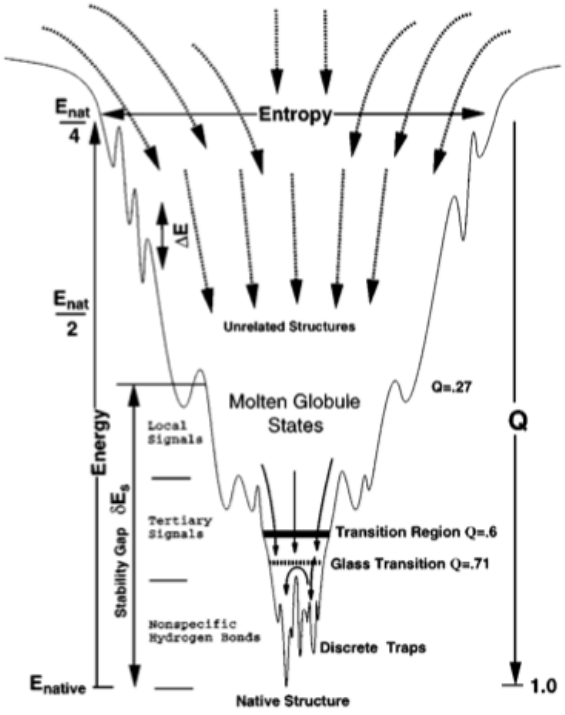


Figure 1.4 Detailed energy landscape highlighting various aspects of folding process such as the local, tertiary contributions. Note the transition region is at Q, the fraction native contacts at 0.6. The native structure is well separated by an energy gap and there is a loss in entropy going from the top to the bottom. The protein gets progressive ordered towards the native conformation. Figure adapted from Onuchic, 1997. [19]

The choice of the reaction coordinate is crucial. Typical reaction coordinates are structure based such as the radius of gyration of the molecule (Rg) or the fraction of native contacts (Q). Simple global order parameters such as number of ordered residues (N) or number of native contacts have also been used. *Pfold*, a kinetically derived reaction coordinate that is defined as a committor function, which is the probability of a given conformation folding into the native state before reaching the unfolded state, has been used. Since it requires exhaustive sampling in molecular dynamics simulations, it is practically a rarely seen reaction coordinate.

1.4.2 Entropic Free Energy Barriers

In the funneled energy landscape, the barriers are mainly entropic in nature. Protein stability is determined by the balance between the native stabilizing energy gained as the protein folds and the free energy associated with chain entropy that is lost with the ordering of the chain. The local imbalances between these two energies directly lead to the observed barriers. The entropy loss is initially large as the chain is beginning to be structured and then decreases slowly as the pre-existing constraints, due to the already formed contacts, limit the freedom of the conformations. As the landscape theory promotes a statistical view of the process, even the top of the barriers, the so called transition states in folding are also an ensemble of structures that are rather on the entropic bottlenecks on the way to reaching the native states compared to high energy activated states of traditional chemical kinetics. When there is always a high

barrier to be crossed to get into the native state well, the process is activated and results in two-state or bimodal behavior.

1.4.3 Downhill folding and other scenarios

A *bonafide* prediction of energy landscape theory is "downhill folding". For some proteins, under native conditions such as physiological temperatures or no denaturants etc. the free energy barriers are very small (~kT) or totally vanishing and the polymer just flows down the funnel like landscape without encountering any barrier to reach the native state. This is called the Type 1 scenario in the landscape theory. In this scenario, the proteins show a downhill behavior under optimal conditions and with any perturbation such as mutations, denaturants or higher temperature that increases the native state energy, the proteins makes a downhill to two-state transition with the appearance of barrier.

According to the theory Type 0 scenario is when under all conditions significant folding free energy barriers simply do not exist. Such proteins are referred to as *global downhill folding* proteins and they have broad thermodynamic folding-unfolding transitions. For such proteins, folding becomes a completely diffusive process in the absence of any free energy barriers. When denaturing stress or perturbations are applied, the free energy minimum (single well) gradually shifts with the population moving towards the unfolded state making it '*one-state*' downhill always. In stabilizing native conditions the single well on the free energy surface corresponds to the native state ensemble. In midpoint conditions, the ensemble as a whole gradually moves towards higher molecular disorder. Under denaturing conditions the well corresponds to the unfolded state ensemble. This gradual shifting is unlike the all-or-none transitions observed for two-state proteins.

Importance of such downhill folding proteins stem from the fact that they enable a complete mapping of the folding-unfolding transition using equilibrium experiments on non-mutated, natural proteins. They provide unique opportunity to completely resolve mechanisms of folding experimentally.

1.5 Characterization of downhill folding proteins

In 2002, Munoz and coworkers first identified and experimentally characterized a natural global downhill folding protein BBL, which is 40 residues all-helical small domain of a large multi-domain complex from *Escherichia coli*[20]. Using a battery of equilibrium techniques like Differential scanning calorimetry (DSC), far-UV Circular dichroism (CD), fluorescence and Förster resonance energy transfer (FRET), they observed disparate thermodynamic behaviors of the protein. Probe dependence and complex observations such as multiple apparent melting temperatures (Tm) from 295K-335K clearly pointed to a non two-state nature of the protein. Using an extensive statistical mechanical model that included the structural features, they demonstrated that the spread in Tm is due to varied partially folded sub-ensembles populating at different temperatures.

In 2006, the structural heterogeneity of such structural subensembles arising due to the small degree of thermodynamic coupling in the protein was characterized and resolved using NMR[21]. Monitoring the chemical shifts of 158 protons as a function of temperature revealed the high structural heterogeneity during the downhill folding of BBL. The analysis of these atomic unfolding curves also produced a broad distribution of midpoint temperatures that was centered at the global midpoint temperature but spanned the whole of the global unfolding transition. Detailed maps of networks of noncovalent interactions stabilizing the native structure and their changes during the unfolding process were obtained from this high resolution atom-by-atom probing of BBL. Thermodynamic coupling between such contacts were revealed that resulted in identification of a small cluster of critical strongly coupled interactions holding the native structure together. Such atom-by-atom analysis has now been extended to few other downhill proteins as well.

In 2012, single molecule characterization of BBL was achieved after many improvements in the single molecule FRET (smFRET) technique[22]. Single molecule measurements were expected to be the ultimate resolution in observing the one state continuous folding of BBL and to unequivocally establish downhill folding of this protein. As a fast folding protein folding in μs timescales BBL posed many practical problems to obtain sufficient photons to derive enough statistics from these experiments. After slowing down the kinetics of the protein with low temperatures and using photo-protection cocktails to employ high intensity lasers for obtaining increased photon counts from the dyes, finally the one state dynamics of BBL was observed. At the single molecule level, the protein only had a single broad population at all the times and shifted gradually from high FRET values in the native conditions to the low FRET values in the unfolded conditions. Various improvements in the analysis procedures along with quantitative modeling were performed to support the interpretation of the data.

Another downhill folding protein, gpW, a 62 residue viral protein with α+β topology was identified in 2007 and has also been extensively characterized using various thermodynamic, kinetic and ensemble approaches[23,24].

1.6 Kinetics of Fast folding Proteins

Of the various developments in the experimental study of protein folding in the past 20 years, ultrafast experimental techniques to measure fast events in the ns-ms time scales are the most important[25]. They have helped identify and characterize many fast processes from the dynamics of elementary events in folding such as the secondary structure formation, loop dynamics etc. to identification of proteins that fold completely within a few μs to hundred of μs. Ultrafast laser temperature jump (T-jump) techniques spawned and spearheaded this line of research in protein folding kinetics. Now an arsenal of techniques has been developed to obtain resolution in the desired ns-ms range. This includes various new equilibrium techniques based on NMR, fluorescence spectroscopy and single molecule fluorescence to different relaxation techniques

like temperature, pressure, pH etc. and microfluidics based ultrafast mixing techniques.

Proteins cannot fold faster than the time taken for their basic components such as helices, sheets or loops to be formed. The dynamics of these elementary processes thus set the limit for the folding rate of a polypeptide chain. Peptide bond rotations take place on the timescales of 1-2ns and peptide bond formation takes ~10ns[26]. Alpha helices form in ~200 ns and beta hairpins take ~1-5 μs. There is a wide variability among the kinetic measurements depending on the sequences of peptides used, positions of the probes used and other factors. But these have been taken as the characteristic times for the formation of the elementary structures[26]. Hydrophobic collapse of the polypeptide chain takes place in 100ns[26]. Loop formation in terms of the chain end-to-end contact formation the takes 10-30ns[27].

The range of timescales observed in fast folding proteins sets upper limits for the folding times between 0.3 and 3 μs at 333 K[26], as derived from the estimate of N/100 μs folding speed limit set by Eaton and coworkers[28]. Eaton and coworkers used a simple linear scaling between protein size and folding rates to come up with the speed limit estimate. Using the number of ordered residues as a reaction coordinate, the time to move down a steep free energy well is simply proportional to the length of the reaction coordinate which is N, the size of the protein. For small proteins, since N is also approximately proportional to the number of native contacts, a simple proportionality also extends to Q, the fraction native contacts as reaction coordinate.

The search for fast folding proteins was driven by two major factors. One motivation was very practical which was the need to identify rapidly folding proteins so that direct comparisons with computer simulations could be carried out. As explained in the previous section, downhill folding, an implication of the energy landscape theory could be observed in proteins that have fast folding kinetics. This was the second major motivation for identifying and characterizing fast folding proteins.

In laser T-jump experiments, Gruebele and coworkers[29] working on engineered variants and mutants of 80 residue alpha helical protein called lambda-repressor found that some of the mutants speeded up the folding enormously to just ~20μs folding times resulting in non-exponential behavior and also introduced a new ~2μs phase in the kinetics. This fast phase becomes the only observable phase when the mutants were further stabilized by using cosolvents. The kinetic argument used was that the fast phase corresponds to depletion of population at the top of the barrier that provides an estimate for the diffusion coefficient to be ~1/(2 μs) at 340 K for the folding of this protein[29-30]. This corresponds to a free energy barrier of ~1.5 RT, very marginal. Splitting of the kinetic relaxation into two phases suggested that for these proteins the free energy surfaces only have shallow barriers and different perturbations could modulate these barriers. The fast phase is diffusive akin to the downhill folding proteins. As the probe dependence in equilibrium experiments observed in the characterization of BBL, they observed kinetic probe dependence in the fluorescent and infrared T-jump

experiments at temperatures below the Tm for the lambda repressor mutants, which was indicative of downhill folding. Gruebele and coworkers have later successfully engineered a lambda repressor variant to fold in a globally downhill manner [31].

1.7 Single molecule experiments

Advances in single molecule spectroscopy – single molecule force spectroscopy and single molecule fluorescence spectroscopy, in particular smFRET and Fluorescence Correlation Spectroscopy have found tremendous application in studying proteins[32]. Over the past 15 years, multiple fundamental observations from mechanical properties of proteins to the heterogeneous subpopulations under equilibrium conditions to one-state folding and characterization of the transition path time during which the protein actually folds have been achieved observing proteins one molecule at a time. Such techniques have truly enabled characterizing previously impossible to study single molecule behaviors of protein molecules. For example, transition path times are uniquely single molecule property and observing transitions of dye labeled individual protein using FRET enabled their measurement[33]. Hopping of protein molecule between multiple hidden intermediates is being revealed only under single molecule force spectroscopy [34]. With the single molecule level characterization of increasingly more proteins, especially proteins with fast folding kinetics, the need for powerful and rigorous methods to interpret and analyze the new data being generated is enormous. The demand for new methods of analysis both to make sense of such complex data as well as to derive the maximum utility from the data being generated from such experiments is high.

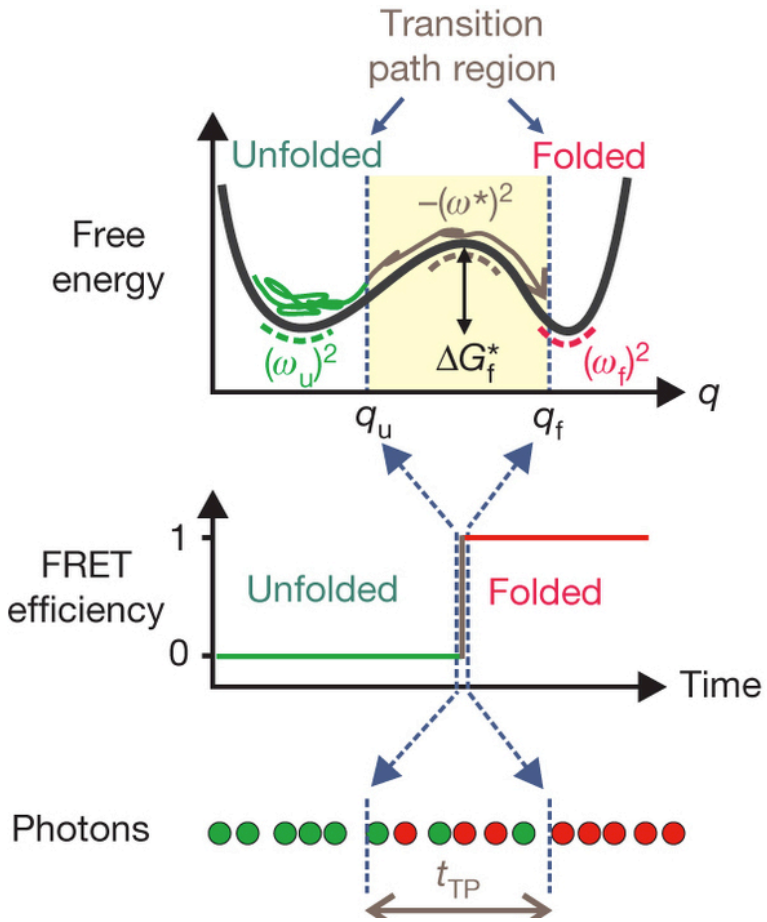


Figure 1.5 Schematic of a folding transition path region with the kinetics given as diffusion on a free energy profile projected over a reaction coordinate q. Photons measured in fluorescence experiments with single molecules showing transitions from folded to unfolded states (emitting green and red photons) were used to calculate the transition path times t_{TP} that are the exact times taken for when the molecule jumps from unfolded well to the folded well fully crossing the region marked as transition path region. Figure adapted from Chung et al. 2013[33b]

1.8 Advances in Molecular Dynamics Simulations

On a parallel front there has been tremendous progress in the computer simulations of protein dynamics. General advances in computing speeds along with development of special purpose hardware and global distributed computing have rendered μs to ms long timescale equilibrium simulations of protein folding possible. A landmark paper in 2011 by Shaw and coworkers demonstrated long all atom molecular dynamics simulation trajectories with multiple folding and unfolding events for 12 different fast folding proteins[35]. Pande and coworkers have achieved aggregate milliseconds dynamics from multiple short trajectories generated from distributed computing[36].

Overall there is a convergence between timescales accessible in all atom molecular dynamics simulations and the experimental data available from characterization of fast folding proteins. This is unprecedented and enables a paradigm shifting iterative collaboration and refinement possible between simulations and experiments.

1.9 Connecting theory, simulations and experiments

Key implication of the energy landscape theory is that the multidimensional protein folding could be projected down into a few suitable reaction coordinates for an enormous gain in the reduction of complexity and be captured as a diffusion process on well-defined coordinates[18]. This feature has been the basis of development of many simple but realistic models of protein folding. It has also been useful for explaining the reasons why such simple models work in the first place.

Simple native centric models such as Go-models that were developed in the 1980s found grounding on the Energy Landscape theory[37]. The minimally frustrated energy landscapes of natural proteins imply that only native interactions or contacts matter. Non-native interactions, the residue level or atomic level contacts that are not observed in the native state do not have big effects on the folding process for most proteins. Go-models take the implication much higher and use a fully funneled energy surface where only the interactions observed in the native structure matter. Such models have taken coarse grained molecular simulations to a higher level with the development of simple native structure based Hamiltonians with models both at residue level and at atomic levels. With simplified Hamiltonians, these coarse grained simulations provide higher scales in terms of both speed and size of the system that could be studied [38]. Simple exact models of protein folding such as the 2D lattice models, 3D lattice models, HP models, perturbed homopolymer model etc.[39,40] and other simple models like the random energy models[41] among others have played tremendous role in elucidating the basic features of the folding process. These have enabled understanding of the sequence effects on folding, effects of interactions, kinetics of folding, thermodynamics of folding etc. Here, we restrict to the class of Ising-like models of folding.

Key discovery by Plaxco et al. in 1998 was that folding rates of simple proteins are determined by their native topology and have a high correlation with "contact order" that was defined as the average sequence separation of the contacting residues[42]. Simply, proteins with higher density of contacts that are local in sequence fold much faster than those with contacts that are farther away in sequence. This makes an intuitive sense but such a simple correlation was indeed startling. This further bolstered the implications of the landscape theory that the native structures and contacts have a larger role to play. The finding had many important implications for simple theoretical models as well suggesting that the knowledge of contact map of the native structure is enough for describing folding kinetics of the proteins.

Another key to the development of simple models of folding is the success of Ising-like statistical mechanical models beginning with Zwanzig et al. and the later expanded and extensive versions by Muñoz et al in explaining important features of protein folding[43]. Zwanzig model used a simple order parameter such as the number of residues in the native conformations and developed analytical solution for the kinetics and thermodynamics of folding that was used to address questions such as Levinthal's paradox and two-state mechanism of folding.

Munoz et al. developed a simple Ising-like model to explain both equilibrium and kinetics experimental measurements on folding of the 16 residue beta hairpin fragment of protein GB1. The model is referred to as Munoz-Eaton model and incorporates structure into the statistical mechanical framework by using the native contacts in the formulation of the partition function. Simplifications such as permitting only a single continuous stretch of native residues and interactions only if all the intervening residues are ordered were made to develop a tractable partition function. This simplification was referred to as single sequence approximation and it reduced the number of configurations to be enumerated from 2^N to $N(N+1)/2$ for a polypeptide of length N residues. The model produced a two welled free energy profile and explained well the observed details such as two-state behavior, single exponential kinetics, and negative activation energy for folding of beta hairpin. Later the approximation was followed by double sequence approximations, including more details into the model such as loop formation entropy etc. to build more detailed models to explain protein folding kinetics [44]. Munoz et al. also used the model in interpreting experimental data for characterizing the downhill folding protein BBL. Further, an elegant procedure to enumerate all the possible 2^N configurations in the model has been developed [45]. This approach has been extensively applied in the study of many proteins[46]. Models similar to the Munoz-Eaton model were simultaneously developed by Alm and Baker[47] and by Galtziskaya and Finkelstein[48] to mainly study the transition state ensemble and compare with mutational experiments and later to calculate rates of folding.

Following the Ising-model approach, Munoz and coworkers later developed a simple free energy surface model that has been extensively applied in the analysis of many different experimental data such as protein kinetics and stabilities as well experiments on many individual proteins [49,50,51]. Though not having the structural resolution this phenomenological model is very simple yet powerful.

The importance of these simple models is in making the connection between protein folding theory, experiments and the more detailed simulations from molecular dynamics that are now becoming more and more accessible.

1.10 Research Objectives

This thesis is organized in two parts focusing on dynamics in proteins and thermodynamics of proteins respectively.

The first part presents an introduction to stochastic kinetic simulations for investigating protein folding kinetics (Chapters 2 & 3). Various applications of stochastic simulations are presented in chapter 2 and 3. Beginning with study of dynamics of elementary structures of proteins – the alpha helix using stochastic simulations, their applications are demonstrated with the dynamics in two-state and downhill folding. Combining stochastic kinetics simulations with simple models of protein folding in unraveling single molecule behavior of proteins and in the analysis of data from single molecule experiments are then presented.

Development of a rigorous procedure using a maximum likelihood approach for deciphering different scenarios of protein folding directly from single molecule photon arrival data follows next. (Chapter 4)

The second part (Chapter 5) presents the statistical analysis of protein structures to derive a key thermodynamic parameter – the entropy of folding and its addition to improve the simple models of folding.

Part I: Stochastic Dynamics in Protein Folding

2 Stochastic Kinetic Simulations and Simple Models of Folding

"*protein is .. the kicking and screaming stochastic molecule that we infer must exist in the solution* " – *Gregorio Weber*, *biophysicist*

Proteins are inherently dynamic systems with their dynamic motions spanning multiple orders of magnitude from picoseconds to minutes. Computer simulations of protein dynamics have been performed for decades beginning from the early pioneering works of Michael Levitt, Arieh Warshel and Martin Karplus[52], who received the Nobel Prize in Chemistry in 2013 for their seminal contributions. A plethora of methods spanning a wide spectrum of approximations have been developed for simulating protein folding and dynamics. In fully detailed molecular dynamics (MD) simulations, the motions of each of the atom in the protein molecule along with solvent molecules are integrated using very short time steps (fs) with Newton's law of motions using appropriately developed force fields. In coarse-grained simulations protein molecules are represented at different coarse levels as simple connected beads, Cα representations, united atom models etc. and the simulations are performed on-lattice or off-lattice using appropriate Hamiltonians. Simulations have offered many powerful insights into the nature of folding processes.

All atom molecular dynamics simulations that use classical equations with empirically parameterized force fields to simulate protein motions are highly realistic and attempt to model the physical interactions resulting in the dynamics of the protein molecule rather than any implicit assumptions about the system. As both the protein and solvent molecules are represented at atomic level details with the energy terms describing changes in energies due to various non-covalent short and long range interactions such as bond stretching, angular motions, torsional rotations, electrostatics etc. the model simulates the folding process more accurately. In principle, MD simulations using such physics based realistic energy functions should provide all the information about the protein folding process that one could care about. But such level of details comes at an enormous computational cost, making it prohibitive or impossible for simulating larger systems relative to more coarse-grained approaches. Moreover, the energy functions and the force fields are still approximations and still require lot of refinement.

There has been tremendous progress in the MD approach to simulate and study protein folding over the past few years[53]. The advent of special purpose machines built purposely for MD simulations and the advances of MD methods using distributed computing as well the high scalability in the algorithms and MD codes along with the speed improvements in general computing hardware have ushered a new era in the all-atom MD folding simulations. Shaw and co-workers demonstrated MD simulations with multiple reversible folding events of 12 different proteins having folding times ranging from us to seconds using

'ANTON' supercomputer[54,55]. Pande and coworkers, with their distributed Folding@Home, have reached millisecond scales of simulations[36]. Even with commodity hardware and GPU based clusters, MD simulations of folding have become more accessible to many research groups. Yet, so far only small (<100 residues) and fast folding proteins are amenable for such full studies with all-atom MD. For studying larger systems coarse-grained approaches are still the usual and available recourse.

More fundamentally though, for a protein having a free energy barrier, folding is a rare event process i.e. the system spends most of the time residing in one of the equilibrium states and state to state transitions are rare and happen very quickly when they really do. The transition path time i.e. the time taken for the protein to cross from unfolded state to the native state is on the order of μs compared to the average longer residence times. For example, for the Fip 35 WW domain protein the folding time is 10μs compared to the transition path time for 0.4 μs as observed in recent simulations[53]. The implication is that focused sampling of the transitions between the different states potentially improves by orders of magnitudes the sampling and information on the actual folding process for a given investment on the computational resources.

Choice of simulation techniques to be employed for studying a particular system depends on multiple tradeoffs – between the computational speed, time available, amount of details needed, resolution level and comparison to the real system measurements. MD though very powerful and offering the highest resolution is a time consuming method and not necessarily the right choice depending on the contexts of the questions being asked.

For probing dynamics of chemical systems or individual molecules where the kinetic phenomena and the discreteness of the molecules and events are of interest rather than spatial and structural resolution, stochastic kinetic simulations are a very appropriate and useful technique. The difference between the molecular dynamics and stochastic kinetic approaches is shown in Figure 2.1.

2.1 Master Equation Kinetics

Basically, a master equation precisely describes the time evolution of a population of species in a well-mixed chemical system or of the discrete state probabilities of single molecules. It is a set of first order differential equations concisely represented. Master equation approach has become ubiquitous in many areas of physics, chemistry and has been increasingly applied in biology. Master equations are also connected with and isomorphic to other standard stochastic models like the Fokker-Planck equation and the Langevin dynamics equations. Kinetics for the probability or fractional population, $P_i(t)$ for the i^{th} conformation ($i = 1,\ldots \Omega$) is determined by:

$$\frac{dP_i(t)}{dt} = \sum_{i \neq j}^{\Omega} [k_{ji} P_j(t) - k_{ij} P_i(t)] \qquad (2.1)$$

In Equation 1, k_{ji} and k_{ij} are the rates of transitions between state i and j. The rate of change of the probabilities depends only on the current state of the system implying a memoryless, Marokovian behavior. Models based on Eq. 2.1 have been used to describe many stochastic phenomena such as single molecule behavior, reaction kinetics, diffusion systems, protein folding and biological networks etc. Deterministic mass-action kinetics could in fact be considered to be mean-field approximations of the above Eq 2.1.

Equation 2.1 could be cast in the form of a vector-matrix formalism as:

$$\frac{dP(t)}{dt} = \mathrm{K}P(t) \qquad (2.2)$$

where K is the rate matrix. The non-diagonal matrix element of K, K_{ij} is the rate constant for transition j -> i ($i \neq j$) and the diagonal element K_{jj} is the rate constant for escape from a given microstate j. The column sums $\sum_i K_{ij} = 0$ for all j which is simply by the conservation of probabilities and from this $K_{jj} = -\sum_{i \neq j} K_{ij}$. As the equations are all coupled first-order differentials, the formal solution of Eq 2.2 at t=T being the final time of interest, knowing P(t=0) is given as:

$$P(t) = \exp(\mathrm{K}t)P(0) \qquad (2.3)$$

where exp(Kt) is a propagator of the system that could be solved using the eigenvector-eigenvalue decomposition as:

$$\exp(\mathrm{K}t) = U \begin{pmatrix} \exp(\lambda_1 t) & & & \\ & \exp(\lambda_2 t) & & \\ & & . & \\ & & & \exp(\lambda_N t) \end{pmatrix} U^{-1} \qquad (2.4)$$

$$P(t) = U\,diag(e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t} \ldots., e^{\lambda_N t}) U^{-1} P(0) \qquad (2.5)$$

For a Markovian, ergodic system with rate matrix K, there is always a unique eigenvalue $\lambda_1 = 0$ and all other eigenvalues must have a strictly real part. The eigenvector v_1 corresponding to $\lambda_1 = 0$ must only have non-negative components as their values are proportional to the probabilities of the macrostates in equilibrium (steady state probabilities).

The elements of the rate matrix are always assumed to satisfy the condition of detailed balance (Eq 2.5), which implies that the eigenvalues of the matrix K are all real.

$$P_{ij}^{eq} k_{ij} = P_{ji}^{eq} k_{ji} \qquad (2.6)$$

In principle, from the solution to the master equation any experimental quantity such as fluorescence intensity or signal could be obtained as a weighted average of the probabilities of the macrostates.

Solution to the master equation as Eq 2.3 is exact and yields a set of 2^N eigenvalues each with associated eigenvector. For smaller number of states, analytical solutions to the master equations are easy. But as N gets larger the solution is non-trivial as the matrix exponentials are impractical to obtain and even numerical solutions become intractable. Also to be noted is that the rate matrix K is often a sparse matrix from the fact that the number of states accessible from (connected with) a given state is generally much smaller compared to the total number of states, especially for larger N. As an alternative to the analytical/numerical solution approach, stochastic simulations that literally are numerical realizations of P(t) versus t, are the way to compute averages and correlation functions for the system of interest by directly computing the state transitions of the system over time. The approach doesn't yield a numerical solution to the master equation which will be the probability density function but a random sample of P(t).

2.2 Stochastic Simulation Algorithm

Over past decades various simulation algorithms have been developed to numerically simulate evolution of system governed by master equation kinetics. Dan Gillespie in 1977[56,57] was the first to develop the direct stochastic simulation algorithm (SSA) for simulating the time evolution of a chemical system as numerical realizations.

2.2.1 Gillespie Algorithm

Let $X_i(t)$ denote the number of species S in the system at time t. For the state vector $X(t) = (X_1(t), X_2(t), X_3(t)...., X_n(t))$, the goal is to estimate it at time t, given the state of the system at $X(t_0)=x_0$ at initial time t_0. Gillespie realized that the key to generating simulated trajectories of X(t) is not the function $P(x,t|x_0,t_0)$ but rather a new probability function $p(j,\tau|x,t)$ that is defined as follows:

$p(j,\tau|x,t)$ = the probability, given X(t) = x, that the next reaction in the system will occur in the infinitesimal time interval $[t+\tau, t+\tau+d\tau)$ and will be R_j (2.7)

This is the joint probability density function of two random variables, time to the next reaction (τ) and the type of the reaction (j) given that the system is in state **x.** The exact formula for the above equation is given as:

$$p(j,\tau \mid x,t) = a_j(x)\exp(-a_0(x)\tau) \qquad (2.8)$$

$$a_0(x) = \sum a_j(x) \qquad (2.9)$$

$a_j(x)$ is a propensity function which gives the probability that a single reaction Rj will happen in a time interval $d\tau$. $a_0(x)$ is the sum over the propensity for each of the reactions in the system.

Equation 2.8 is the basis for the SSA approach and it implies τ is an exponential random variable with mean $1/a_0(x)$ and j is another independent random variable with probability $a_j(x)/a_0(x)$.

The easiest implementation of the SSA is called direct method that is as follows: Draw two random numbers r_1 and r_2 from the uniform distribution and compute $\tau = 1/a_0(x)\ \ln(1/r1)$ and j = smallest integer satisfying the condition $\Sigma a_j(x) > r_2.a_0(x)$. Then the following steps are performed:

1. Initialize the time t = t0 and the state of the system x = x0
2. For system in state x, evaluate all the $a_j(x)$ and their sum $a_0(x)$
3. Generate values for τ and j (using above random number criteria)
4. Do the next reaction by moving from t<- t+ τ and the state vector x
5. Record the current state x(t),t. Repeat from 2 as desired or end simulation

The direct SSA algorithm has been applied widely but it is slow when the numbers of the reactions become large or the rates become very fast so that a large number of realizations has to be performed. Many alternative methods such as τ-leaping and approximations on the algorithm have been developed to speed up the stochastic simulations procedure. As could be noted from above, both the time of the next reaction and its kind are computed on the basis of stochastic reaction constant (propensity constant) that by definition determines all the other aspects of the SSA. But this constant itself is an unknown and is only guessed in an approximate manner starting from the deterministic rate constants.

2.3 Recipe for probabilistic kinetic simulations for protein folding transitions

For reactions systems or individual molecules with discrete states, when the rates of the reactions or interstate transition rates are already known either from experimental measurements or theoretical estimates, we use a variant of the stochastic algorithm to obtain realizations of state trajectories. Our recipe is a distinctive variant of the above direct Gillespie SSA that uses a constant predetermined time step Δt rather than updating the times from exponential distribution at each step. Key assumption of our probabilistic method is that at any time step Δt at most one and only one reaction may occur. The length of Δt (i.e. time step) depends on the number and speed of the chemical reactions or transitions and is chosen so as as to render small transition probabilities ensuring utmost single transitions within each step.

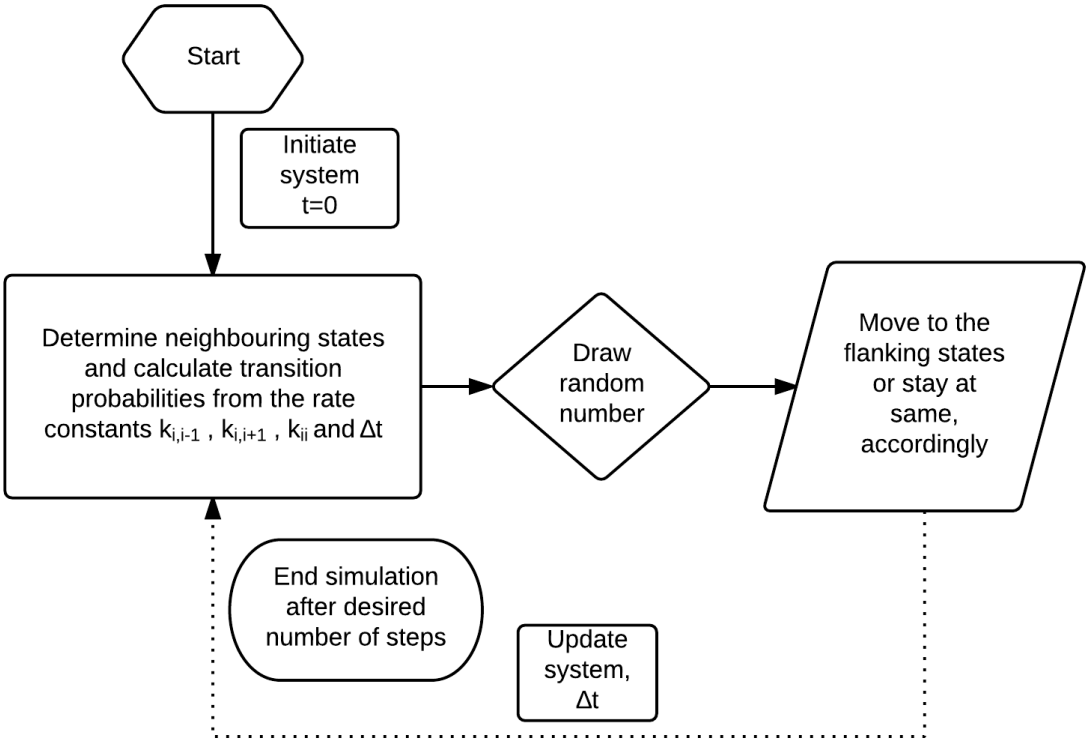


Figure 2.1 Flowchart describing the probabilistic kinetic simulation

For a given Δt, the probability constant for a particular reaction R_μ is given as:

$$P_\mu = k_\mu . \Delta t \qquad (2.10)$$

where k_μ is the rate constant of the reaction R_μ. Therefore the probability of choosing and performing a particular reaction R_μ in the time interval t+ Δt out of M total reactions is given as

$$P(\mu) = \frac{1}{M} P(R_\mu) \qquad (2.11)$$

And the probability of performing no reactions in the same interval is given as:

$$P_o = 1 - \sum_{\upsilon=1}^{M} \frac{1}{M} P(R_\upsilon) \qquad (2.12)$$

Algorithm is given as follows:

1. Initialize the system at t=0 at state x = x0
2. Take a step having a constant discrete time Δt
3. Obtain P_μ , the probability constant for all the possible jumps from state x(t-1)
4. Generate random number r_1
5. Depending on the state x-1 and r_1, make the jump (reaction) or not.
6. Record the current state x(t). Repeat from 2 as desired.

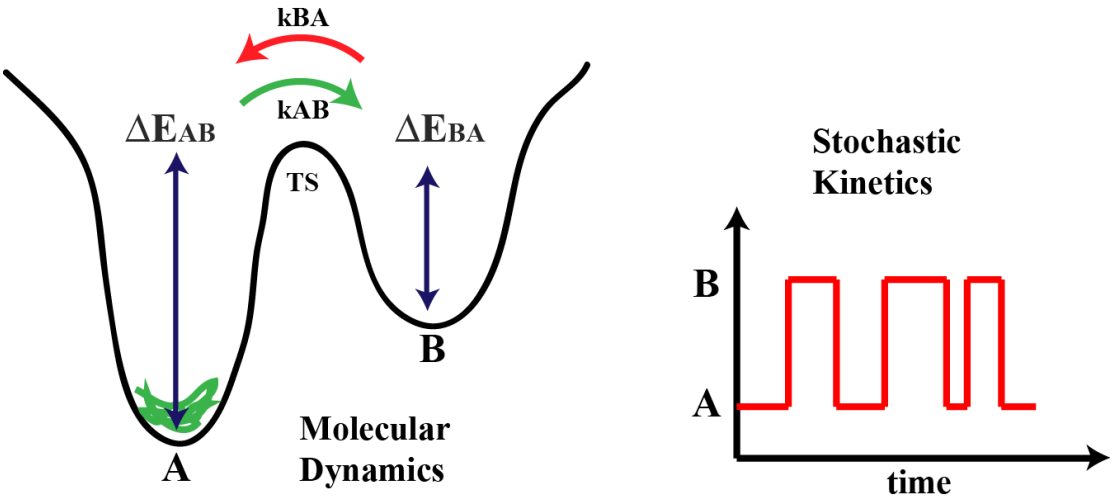


Figure 2.2 Schematic showing the essential difference between the approaches of Molecular Dynamics and Stochastic Kinetic Simulations

The following examples of application of the stochastic kinetic algorithm for protein folding are presented:

Helix-coil Kinetics
Two-state Kinetics (stochastic simulations of simple two-state model)
Stochastic Kinetics of Downhill proteins (on simple harmonic well)

2.4 Stochastic Simulations to analyze Helix-Coil Kinetics

This work has been published as Waltzing α-helices. Victor Munoz and Ravishankar Ramanathan, PNAS Vol. 106, 1299-1300 (2009)

α-helices comprise the predominant type (~30%) of the secondary structure elements found in proteins. Their existence together with that of β sheets were first theoretically predicted by Linus Pauling based on the hydrogen-bonding patterns observed in proteins. Besides their abundance, α-helices have another important feature that has made them objects of intense research over many decades – they are autonomously folding entities. Despite being simpler secondary structure elements, the formation of α-helices carries all the complexities observed in full protein folding. The conformational behavior displayed by α-helices is wider and more complex than one expects of these simpler units of proteins. α-helix formation involves an array of physical interactions that are all invoked in full protein folding and similar forces and principles govern both these processes. Energetics of interactions including hydrogen-bonding, hydrophobic effects, electrostatics, dipole-dipole interactions, van der Waals interactions along with the entropic forces are all invoked in α-helix formation. Thus α-helices serve as excellent model systems for understanding and elucidating the fundamentals of protein folding. Extensive studies over the past 60 years have established well the thermodynamics of the process whereas the kinetics of it, a harder problem has also been fully unraveled in the past 20 years with the development of laser induced ultrafast

nanosecond temperature jump techniques and with the developments in single molecule fluorescence spectroscopy.

The defining interactions in the α-helix are i,i+4 hydrogen bonds in the main chain of the polypeptide between carbonyls of residue i and the amide protons of residue i+4. Along with these hydrogen bonds, dense packing of the backbone with 3.6 residues per turn results in favorable interactions like stabilizing van der Waals forces. The dipolar peptide bonds get oriented in an energetically favorable direction parallel to that of the helix resulting in a macro-dipole with a net positive partial charge in the N-terminus and net negative partial charge in the C-terminus. All the side chains orient outwards and tangentially to the backbone. The staggering avoids steric clashes and facilitates side chain-side chain interactions between i,i+4 and i,i+5 residues. All the amino acids are incorporated regularly in the helix except proline due to its cyclic imide group produces a kink in the chain and lacks the hydrogen bond donor and thus breaks the regular α-helix. Glycine on the other hand with lack of side chain and steric freedom incurs a heavy entropic cost for getting ordered within the helix and hence is less favored.

Theoretical underpinnings to the helix formation, referred to as helix-coil transition was formulated in the 1950s. α-helix formation was described essentially as a nucleation-elongation process that involves the entropically unfavorable initial step of nucleation wherein four consecutive residues need to be simultaneously structured to 'nucleate' the helix, which could then propagate bi-directionally with the addition of further residues. The elongation process is relatively easier with a net enthalpic gain arising from the many interactions that have been well catalogued. Long helix forming homo-polymers were the initial model systems used to study α-helix formation and several pioneering experiments helped to characterize the process. With the identification of independently folding α-helical fragment from a natural protein and the subsequent development of design principles to produce short helical peptides [58] ensued extensive experimental characterization of factors leading to α-helix stability. Parameters in the helix-coil theory namely the nucleation parameter (σ) and elongation parameter (s) were well established to the level of individual amino acids preferences, which were incorporated into the Zimm-Bragg model to predict sequence dependent properties. Empirical data obtained from hundreds of such designed and natural α-helices were incorporated into the AGADIR force field [59], which formed a bedrock with real power for predicting helix content of peptides under different conditions of temperature, pH and ionic strengths. AGADIR demonstrated that the general principles of α-helix stability have been understood and that the theory was well established.

The most widely used techniques to study helix-coil dynamics are relaxation experiments. Seminal connections between the kinetics and thermodynamics of the process described with the helix-coil kinetic theory was made earlier on by Schwarz[60] and the results of early studies on long polypeptides summarized by Grunewald et al[61]. Schwarz's analytical formulation proposed the relationship between σ, s, k_f the rate constant for helix propagation and τ^*, the mean relaxation time.

$$\tau^* = 1/k_f(4\sigma + (s-1)^2) \qquad (2.13)$$

The barrier for helix nucleation is larger than any of those for the local changes in the helix and so the gain or loss of helical sequences is expected to be the slowest process in helix kinetics. At the midpoint of the transition, $s \sim 1$ and $\tau^* = 1/(4\sigma k_f)$ and is at its maximum estimated at $\sim 0.1\mu s$ and $k_f \sim 10^{10}/s$ and σ estimated around 10^{-4}. The sharper the transition (smaller the σ) the slower the transition takes place. The elementary step of growing the helix is expected to be a diffusion-controlled formation of hydrogen bond.

Helix coil kinetics was among one of the first phenomena to be measured using the ultrafast folding techniques notably T-jump techniques that came into foray in the last two decades. From these studies, relaxation rates for helix coil transitions were measured to be in the hundreds of nanoseconds [62] [63] [64] and the elementary rates of rotations of single peptide bonds from coil to helical and vice versa estimated to be 1-4 ns per residue.

The nucleation barrier produces the slowest relaxation time in the helix-coil process both during formation and melting because of the fact that in the equilibrium studies done near midpoint temperatures, the total height of this barrier is relative to both the coil states and the formed helical states. When this barrier is large, the process is a complete two-state like all-or-none transition and this barrier crossing is the only observed process. Whereas when it is smaller, changes in helix stability lead to a distribution of both numbers and lengths of helices. Because of the barrier, helix nucleation in a completely unfolded protein chain or coil nucleation in with the helical sequence of the protein are both rare events. Since adding or removing residues to existing helices are rapid compared to crossing this nucleation barrier, helical peptide sequences reequilibrate to the new distribution of lengths at timescales much faster than their equilibriation with coils. These helix propagating-shortening events should take more than the 1-4 ns of per residue rotations but are distinctively faster than the motion over the barriers and equilibration with coils. However, these fast processes weren't observed in the T-jump experiments (neither infrared nor fluorescence).

With subsequent combination of T-jump experiments and single residue level detection by isotope labeling and using infrared, complexity of helix-coil kinetics became more apparent[65]. Apparent relaxation times were found to be dependent on the magnitude of perturbations including the size of the temperature jumps and in addition stretched exponentials were observed for single peptides irrespective of position of labels. The apparent kinetic complexity at residue level warranted novel statistical mechanical models of helix-coil kinetics that were further developed to address it[66]. This detailed theoretical analysis synthesized both equilibrium and kinetic details of the helix-coil transition to provide a total quantitative understanding of the mechanisms of α-helix formation. But, the diffusive motions of propagating-shortening pockets of helices in peptides still remained unsolved and to be observed in any experiments. Meanwhile, molecular dynamics simulation studies of poly-

alanines [67] also suggested such diffusive conformation search happening in helices though results from these studies weren't totally reconciling with the established helix-coil nucleation theory.

Earlier, Lapidus et al[68] used tryptophan fluorescence triplet quenching to study end-to-end contact formation and global dynamics in α-helices. They measured the time it took for the quenching of triplet states resulting from nanosecond laser excitation of tryptophan residues attached to one end of 22-residue long polyalanine peptide upon formation of contacts with cyclic disulfide attached to the other end of the helix. Analysis of the fluorescence decay of the triplet population yielded diffusion limited end-to-end contact formation rates of this peptide along with the helix->coil and coil->helix rates. The helix->coil timescales measured were consistent with those measured from the temperature jump experiments whereas the end-to-end contact formation rates were measured to be 1.1×10^7 per s in the coil state. This work measured the global dynamics of contact formation in helical sequences.

Firez et al,[69] used ultrafast contact formation to address the question of local dynamics in helices and observed intra-helical diffusive motions under equilibrium conditions using a technique referred to as triplet-triplet energy transfer (TTET) that probes local dynamics. By attaching donor xanthone and acceptor naphthalene at various i,i+6 positions in a 21 residue polyalanine peptide and measuring the energy transfer the local dynamics was probed

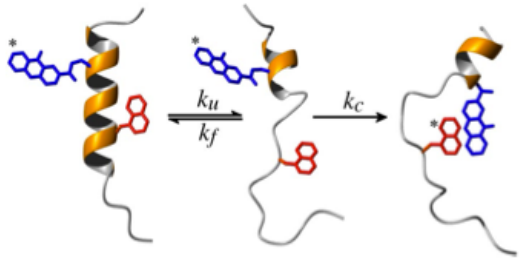


Figure 2.3 Schematic to show the donor and acceptor dyes attached at i, i+6 positions that are off register and on the opposite faces of helix to monitor local dynamics. (adapted from Firez et al)

Contact forms when the intervening residues transition to coil states thus bringing the dyes close enough in space for TTET to occur (schematic shown in Figure 2.2). As i, i+6 positions are out of register in an α-helix, a contact between these moieties could form only when the intermediary residues unfold and become coil like. The rate of contact formation is thus coupled to the local helical motions at equilibrium and the measurements report transient coil-like gaps between the attached probes in the strongly helical sequence. The rates of elementary events of helix elongation and helix shrinking were extracted to be 40-60ns from the experimental measurements based on the microscopic interpretation that the contact formation rate acts as reporter of the local helix motions. This was the first time the local dynamics in helices were experimentally measured.

With the solid theoretical foundations and models capable of explaining complex helix kinetics, as a way to independently corroborate these experimental results

we performed stochastic kinetic simulations. Stochastic kinetic simulations offer a powerful way to obtain mechanistic insights into the local dynamics and motions observed in α-helix and in analysis of such experimental data. Here we use a simple model[70] developed based on helix-coil theory to perform stochastic simulations and obtain single molecule trajectories of local helix motions. From our simulations, it will be evident that the approach is powerful and yields direct insights into the dynamics of the system.

2.4.1 Simulation of Local α-helix dynamics with stochastic kinetic model

In the model, a peptide bond could be either helical or coil and rotations between these two-states are considered to be the elementary kinetic steps.

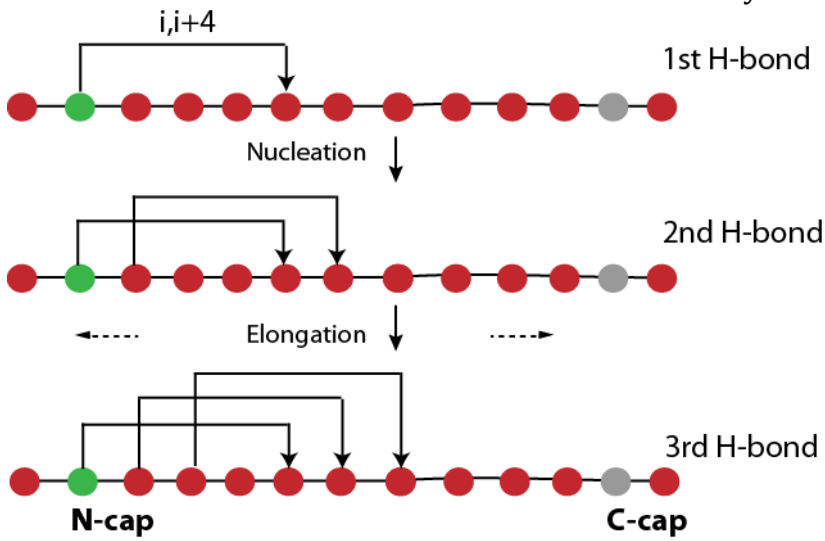


Figure 2.4 Schematic showing the helix nucleation and elongation Helix is nucleated with the formation of a i,i+4 hydrogen bond which is the slowest process. After nucleation, the helix could simply propagate by adding more hydrogen bonds on either direction.

2.4.2 Description of the nucleation-elongation based stochastic model

The original model [70] has been formulated as a more sophisticated version of the well-established nucleation-elongation theory used to interpret T-jump experiments, with the key introduction of sequence dependence and double-sequence approximation. Sequence dependence is incorporated by using the set of parameters for the helix-coil transition in equilibrium taken from AGADIR that was itself parameterized from extensive compilation of experimental studies. By incorporating different amino acid propensities and positional preferences (N-capping, C-capping in helices) the model adds lots of details. The double sequence approximation simply is allowing two non-overlapping segments of helices to be present simultaneously in a single molecule, allowing helices to break from the middle and the merging of short fragments to form longer ones. This is a crucial factor in the model as it enabled the study of mechanistic formation of helices and the independent local dynamics of the segments in helices. The model explained well complex helix-coil kinetic behavior as measured by T-jump experiments in 13C labeled peptides and the physical origin of non-exponential time courses and observations of different relaxation times for various regions of the peptide and for various magnitudes of perturbations.

Peptide bonds adopt two-states - either helical (h) or coil (c) conformations. When the flanking peptide units (ϕ_{i+1}, ψ_i) adopt torsional angles in the helical region of Ramachandran plot the residues are marked helical, coil otherwise. Coil state is taken as the reference and hence has statistical weight w_c =1. Fixing a pair of dihedrals to the helical region versus the whole dihedral space it could occupy incurs a conformational entropic cost and the more number of residues ordered, higher the entropic cost. This makes the intrinsic statistical weight for helix state h to be < 1 (h_{in} = exp(ΔS/R) < 1) where ΔS is the entropic cost of helical ordering (ΔS = S^H - S^{coil}) and R is the gas constant. Consecutive ordering of five residues (intrinsic weight given by $(h_{in})^5$) and the formation of i,i+4 H-bond interactions and other favorable interactions causes the nucleation of the helix, which is the rate limiting step. Further elongation of the helix only need the fixing of the dihedral angles with addition of H-bonds resulting in net gain of stabilizing energy from the backbone interactions. Statistical weight for adding another h via elongation is given by products of h_{in} and h_{bb}= exp(-ΔG/RT). Using parameters from AGADIR for adding in sequence dependence and amino acid specific properties, the intrinsic helix weight is given as $h_{in} = \exp(-(\Delta G_{in,i} + \Delta G_{in,i+1})/(2RT))$. The statistical weight of a helical segment with j peptide bonds in h is given as:

$n(h_{in})^j c$ for j<5 and $n(h_{in})^j (h_{bb})^{j-4} c$ for j>=5

where n,c are the weights for N and C terminal caps respectively, also parameterized based on AGADIR values. For a peptide of N bonds (N+1 residues), there are 2^N conformations or species possible.

Using the double sequence approximation introduced above, the total number of species drops to $\left(\frac{N+1}{2m}\right)$ with 1, 2 or no helical segments (m=2 for double sequence approximation). For a 22 residue peptide studied here this means, from 2^{20} the number of species becomes 6196, a drastic and useful reduction. The partition function in the double sequence approximation is given by:

$$Q = 1 + (\sum_{i=1}^{n} \sum_{j=1}^{n-i+1} w_{ij} (1 + \sum_{p=1}^{n-i-j} \sum_{q=i+j+1}^{n-p+1} w_{pq})) \qquad (2.14)$$

where $w_{ij} = \exp(-\Delta G_{ij} / RT)$ and $w_{pq} = \exp(-\Delta G_{pq} / RT)$ are statistical weights of helical segments of i and p number of peptide units that start at positions j and q of the molecule. The probability of any conformation is then given by $w_{ij} w_{pq} / Q$ where for i=0 and p=0, w_{ij} and w_{pq} are set to 1. Of note to our stochastic kinetic simulations are that since only realizations of trajectories are needed and the enumeration of the thermodynamic states is not required, the double sequence approximation is not invoked. The nucleation-elongation model is used as is with the amino acid specific interaction parameters obtained from AGADIR. The full local dynamics of helix with coil to helix flips and vice versa are thus allowed in the simulations.

2.4.3 Kinetics in the model

The elementary steps in the model are bond dihedral angle rotations from coil to helix (on rate) and vice versa (off rate). The on rate is expressed as $k_{on} = k_o h_{in}$ where k_o is the pre-exponential factor defining the rate of peptide bond rotations. k_o is an adjustable parameter and sets the absolute time scales of dynamics of the helix-coil transitions. The off rates are obtained by applying the principle of detailed balance and are given by $k_{off} = k_o h_{in} w_{j+1}/w_j$ where w_{j+1} and w_j are the statistical weights of the final and initial conformation that differ by a single helical peptide bond. A value of $k_o = 2.5 \times 10^8\ s^{-1}$ is used for all calculations. The master equation is then derived from these elementary rates and the kinetic rule that each conformation is connected to any other conformations that are only accessible by single peptide bond rotations. The master equation is expressed in a numerical form and is typically a sparse matrix given the above kinetic rule. This matrix is then solved numerically using standard methods for stiff problems.

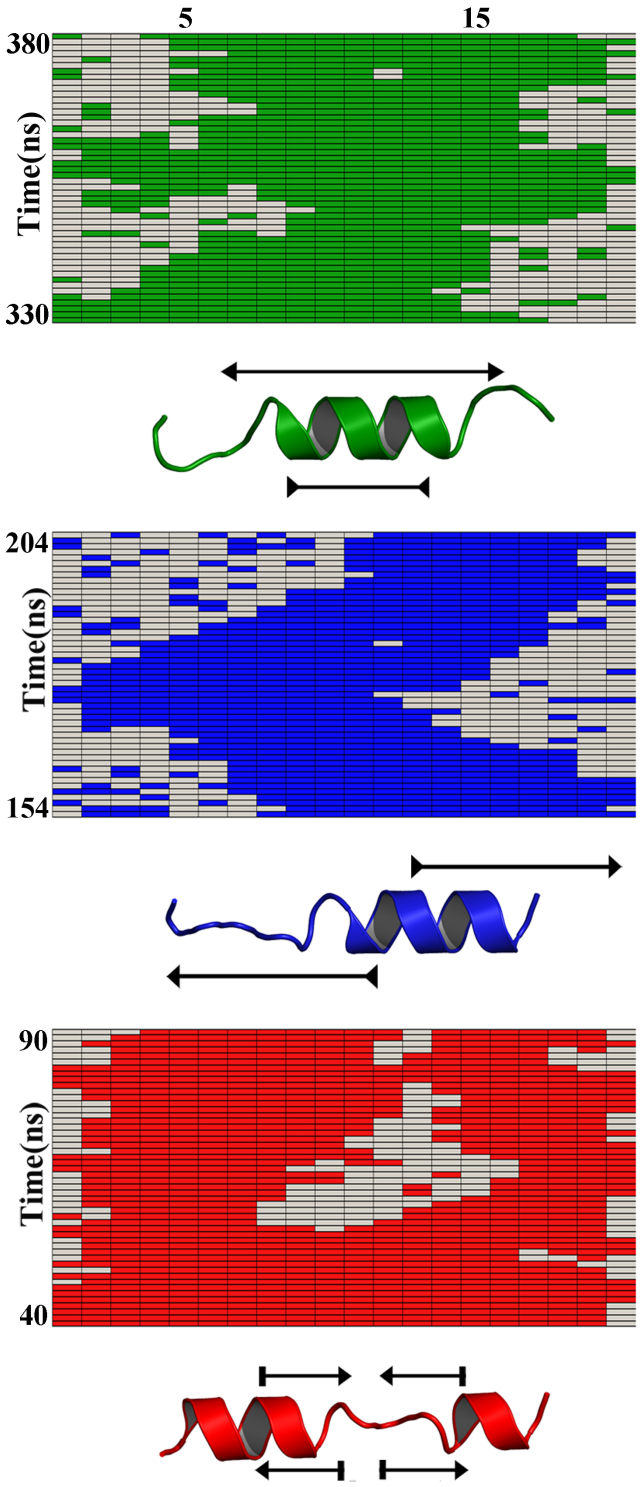


Figure 2.5 Stochastic kinetic simulations showing different local movements in helix of a 20 -residue peptide according to nucleation-elongation theory. Rectangles show 50-ns segments of the stochastic simulation showcasing the 3 basic helix motions: green shows stretching-shrinking; blue shows sliding; red shows splitting-merging. Simulations are performed with single residue rotation rates estimated from T-jump and are nicely consistent with Fierz et al. results (adapted from Ramanathan & Munoz[71])

2.5 Stochastic two-state Kinetics

2.5.1 Background on two-state kinetics

Two-state kinetics is observed in many biological processes. In protein folding since the observations of Fersht and coworkers that small, single domain proteins fold cooperatively[72], two-state model has been applied very readily to conclude that many such proteins do possess two-state behavior. This has led to the situation that most protein folding kinetics experiments are interpreted with this simple chemical model by default and if only there is any difficulty in fitting measured data to the model, alternative models such as presence of intermediates are invoked. That too typically after the failure of trials using different baselines and interpretations to fit the data into this paradigm.

A two-state transition as the name implies is an all-or-none transition. There are only 2 states possible or more precisely observable at all the times for such systems. The system is highly cooperative with concerted transition happening from one state to the other as favored by the thermodynamics of the process depending on the conditions. This is reflected as a sigmoidal curve when the population is followed with a suitable experimental signal such as fluorescence. The implication of the model is that the measured signals could be represented as a combination of those of the individual signals of each of the states. Thus, for two-state proteins, at any time the measured signal has a fraction from the folded and the unfolded states, the individual signals of which are represented as baselines. For an unfolding transition, a pre-transition will mainly correspond to the signal of the folded state and post-transition it will mainly be that of the unfolded state. During the transition, the total signal is a sum of the two signals from both these states.

Two-state models typically lead to single exponential kinetics. This is mainly because of the rapid readjustment of the populations of to a new equilibrium after perturbation during the kinetics measurements. A crucial assumption here that is in fact a precondition for two-state model is that the intra-state transitions are very fast compared to the inter-state transitions. For proteins folding with two-states, this means that there is rapid equilibration within the unfolded and native state ensembles. As the native well is a characteristically a narrow one compared to the typically broad unfolded well this implies that the molecule samples the broad unfolded region rather fast with the conformational transitions inside this well happening much quicker compared to escape from this well.

Another main feature of the two-state model is the transient transition state. This model by definition precludes the possibility of observing any population or ensembles of conformations that are at the top of the barrier separating the native and the unfolded wells. These are highly short lived and are not experimentally accessible. Thus, only indirect inferences have been made about the transition states. Phi-value analysis, a pioneering approach that has led to structural interpretations about the transition state ensembles in two-state

folding proteins in fact measures the effect on stabilities and kinetics relative to the native state.

A battery of tests exists to confirm whether a particular protein indeed folds with a two-state mechanism or not. The criteria for a two-state system are the following:

a) Sigmoidal unfolding transitions on denaturation either thermally or chemically
b) Single exponential behavior in the kinetics under different conditions
c) Thermodynamic parameters identified from equilibrium and kinetics experiments should agree.
d) Probe independence in both kinetics and equilibrium measurements i.e. different probes should all yield similar parameters
e) Identical sensitivity to chemical denaturants in both equilibrium and kinetics; folding and unfolding regimes should be linear as observed in the kinetic chevron plots that are plots of denaturant concentration vs. observed relaxation rates
f) Calorimetric measurements should point to single transitions and satisfy the calorimetric criterion

Importantly, some of these criteria are necessary but not sufficient for a two-state behavior. And deviations from these criteria are only dis-qualifiers but not confirming of any other specific mechanisms. For example, if a non-exponential behavior is observed in the kinetics of a protein folding that definitely disqualifies the protein to be a non two-state system but doesn't necessarily conclude a downhill mechanism of folding. On the other hand, if an exponential kinetics is observed, that by itself doesn't conclude a two-state mechanism that is also the case with observing sigmoidal unfolding transitions in equilibrium measurements. Other folding mechanisms such as downhill folding could also be producing such observations and other tests are needed to conclude any particular mechanism. Calorimetric criterion is an unequivocal test for two-state behavior along with the strict reversibility in these measurements.

At the single molecule level, given sufficient resolution is achievable in the experimental technique two-state behavior is absolutely discernible. An appropriate signal could directly reveal the current state the molecule is in and the stochastic transitions of the molecule between the two-states could be identified. For slow folding proteins, the existence of two peaks in FRET under different stability conditions, unequivocally points to the presence of two-states. And in FRET even under equilibrium conditions the existence of more than one state could be identified. As at single molecule level, even under conditions favoring a particular state, the molecule stochastically jumps to the other state e.g. native conditions in which the folded state is favored, unfolded state are also observed in two-state proteins. Thus, single molecule measurements are the signature resolving techniques for identifying two-state behavior versus alternative mechanisms.

$$U \underset{ku}{\overset{kf}{\Leftrightarrow}} F$$

The transitions between the states are described with the folding and unfolding rate constants k_f and k_u. For the equilibrium of this transition, the equilibration constant is

$$K_{eq} = k_u/k_f \qquad (2.15)$$

In the kinetics measurements, only the relaxation rates are measurable either by thermal or chemical denaturation. The relaxation rate k_{obs} is a sum of both the folding and unfolding rates and is given as

$$k_{obs} = k_f + k_u \qquad (2.16)$$

Typically, the individual rates are obtained by directly solving these equations Eq. 2.15 and Eq. 2.16, which by definition has an implicit assumption of a barrier separating both these states. Of note is that the heights of the barriers are typically not estimated from the above equations.

2.5.2 Stochastic realizations of two-state transitions

Given a two-state folding protein, using our probabilistic stochastic simulation approach, the recipe for which is given in the section 2.2, we could simulate multiple realizations of the transitions between the two-states. From the simulations, various observations about the system such as the equilibrium constant, mean residence times, etc. could be directly inferred. Mapping appropriately signals for each of the states, the net signals could be generated from their summation. This approach, a key ingredient of the two-state model could be used to analyze and model various experiments that follow the signals like fluorescence from the protein. In the following section, the method is illustrated with a simple kinetic simulation.

For a two-state protein folding rapidly, Figure 2.4 shows a sample stochastic kinetic simulation performed according to the recipe given in section 2.3. The forward and reverse rates are taken to be k_f = 3949 s^{-1} and k_u = 1836 s^{-1} and with a short time step Δt = 25 μs, stochastic realizations were obtained. The simulations reveal the dynamics in the molecule with their interstate transition rates defined. By assigning signals to each of the state, different experiments could be modeled as such. For example, if FRET values of 0.8 and 0.55 are assigned for folded and unfolded states, the stochastic trajectories could be converted into FRET trajectories by appropriate mapping. Photon emissions could be simulated using Poisson variables and using the donor and acceptor emissions, FRET could be modeled.

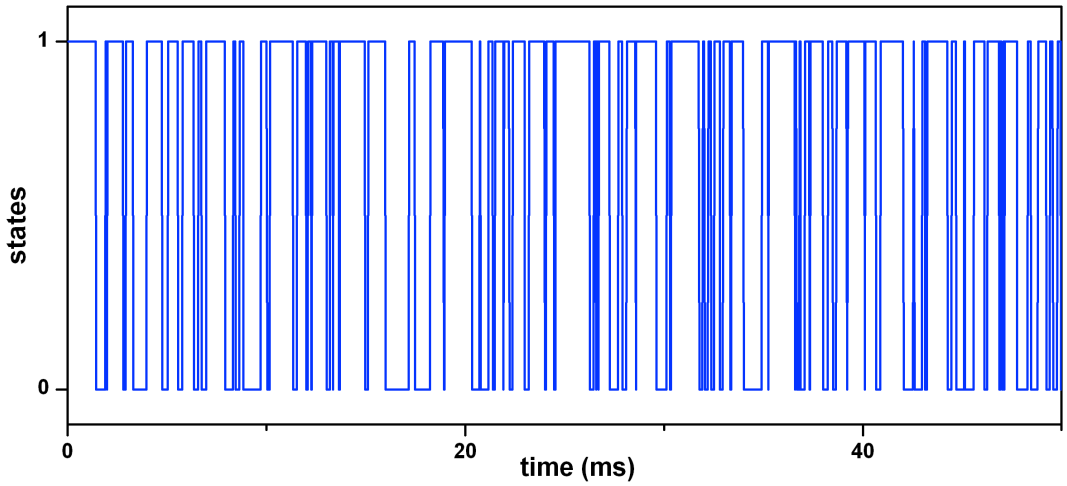


Figure 2.6 Sample stochastic two-state trajectory

2.5.3 Dwell Time Distributions

Kinetics of the system could also be characterized directly from the simulations. In single molecule measurements, kinetics are obtained by calculating the 'dwell

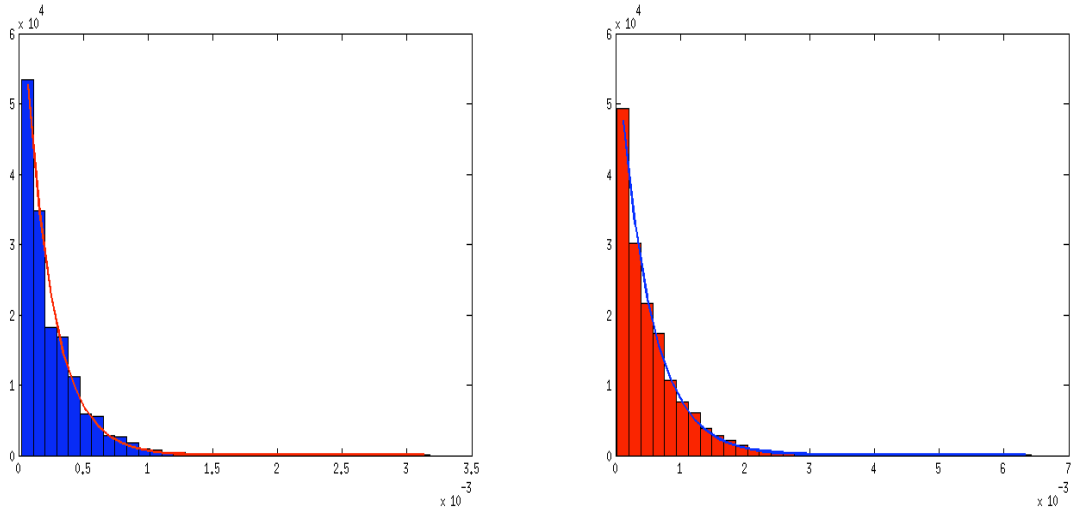


Figure 2.7 Dwell time distributions for the folded (blue) and unfolded (red) states from the stochastic simulations Fitting them to single exponentials provides a way to calculate back the kinetics. When analyzing experimental data, simulating experimental signals such as FRET values offer a way to directly compare the measured and simulated dynamics

times' or 'mean residence times', the average time a molecule spends in particular states. These are directly measured in the single molecule experiments such as FRET or atomic force microscopy. Typically a frequency counting of the times the signal from the molecule corresponds to that of particular states reveals an exponential distribution and fitting it to single exponentials directly provides the relaxation rates from the corresponding states. In Figure 2.5, example dwell time distributions for the folded and the unfolded states calculated from the stochastic simulations of the two-state protein are shown. In this case, as the rates are known and are inputs for the

simulations, the fitted exponentials correspond to these rates. The simulations were done at T = 298K.

2.6 Stochastic Simulations of Downhill Folding Proteins as fluctuations on harmonic well

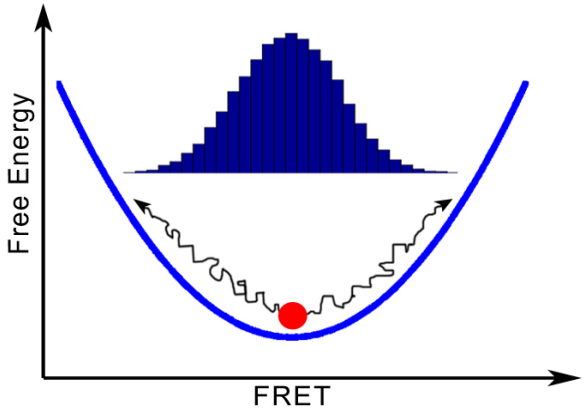


Figure 2.8 Downhill folding as stochastic fluctuation on a harmonic well

Downhill folding proteins, the proteins that fold without any significant barriers (or <1RT) have purely diffusive kinetics since there are no barriers separating the states.

In this case the stochastic kinetics are modeled as diffusive on a simple energy surface of the form F = -RT ln(p), where p is the probability of the states. This probability could also be mapped to a signal such as FRET efficiency values (as shown in the figure) and could be having a shape of an appropriate distribution such as normal or lognormal distributions. The potential is divided into a 100-point space and each of the points is considered to be a discrete state. The molecule jumps between these states depending on the experimental conditions with time dependent probabilities as given below. Here, a harmonic potential energy surface was taken and thus the probabilities are given by a Gaussian distribution with position and width of the distribution being the two main parameters chosen for the model. Kinetics is modeled as a diffusion on this free energy surface and according to the equations derived using matrix formalism as given below [27]. The time dependent probabilities are given by:

$$p(i \to i+1) = \Delta t \frac{1}{2}\left(\frac{p_{i+1}}{p_i} D + D\right)$$

$$p(i \to i-1) = \Delta t \frac{1}{2}\left(\frac{p_{i-1}}{p_i} D + D\right)$$

$$p(i \to i) = 1 - [p(i \to i+1) + p(i \to i-1)]$$

As given there are three possibilities at any given small time step Δt that is chosen such that it guarantees that the probability of jumping to either of the flanking states is always <0.01. The molecule could stay in the current state, could jump to the next state or could jump to the previous state. D is an effective

intramolecular diffusion coefficient that is the key parameter determining the timescales of the process. It is adjusted to reproduce required overall relaxation times as observed in the experiments that are to be simulated.

A stochastic trajectory from a simulation performed using a Gaussian with parameters, mean μ=0.72 and σ=0.08 is shown below in Figure 2.6.

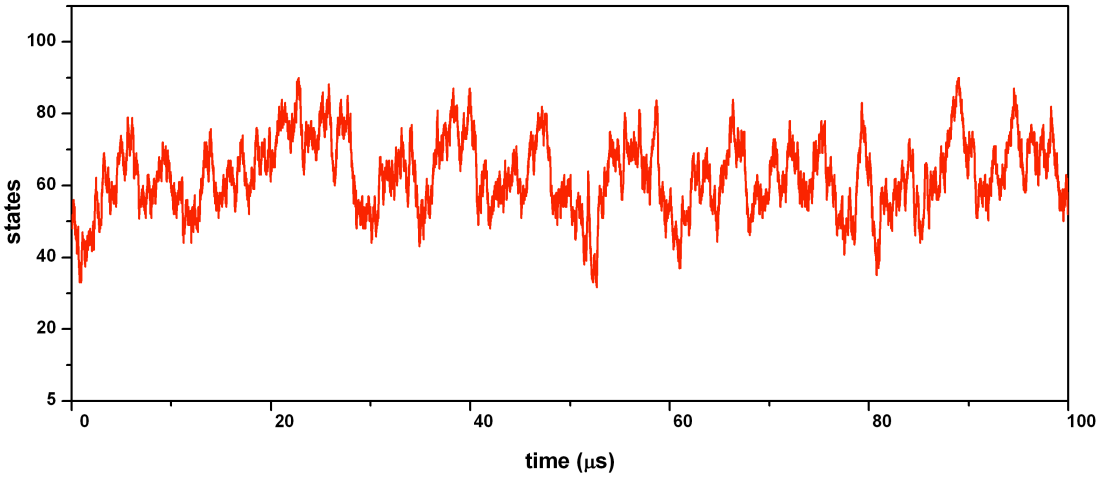


Figure 2.9 Sample stochastic trajectory of downhill folding. The diffusive motions are evident in such simulations.

2.6.1 Kinetics from autocorrelation function

According to fluctuation dissipation theorem [73] the rates of relaxation of system after a small perturbation to equilibrium and the time correlation of spontaneous fluctuations in a system that is not disturbed and at equilibrium e.g. a single molecule being observed under equilibrium are described by the same rate coefficients. Thus, correlation analysis of molecular fluctuations either from single molecule experiments or bulk experiments such as fluorescence correlation spectroscopy or those observed in simulations could be used to provide kinetic information. For analyzing such fluctuations, the central player is the "correlation function". The autocorrelation function of a property called *S*, is defined as

$$\langle S(t)S(t+\tau)\rangle = \lim_{T\to\infty}\frac{1}{T}\int_0^T S(t)S(t+\tau)\,dt$$

Practically, the averaging is done over finite time in experimental measurements as well as in analysis of fluctuations in simulations. Analogous expressions could be defined for cross correlation functions between different properties or signals. The autocorrelation function of a nonconserved property decays typically with a single exponential profile with a characteristic relaxation time or correlation time of the property. Thus, for a protein folding the autocorrelation of any measured signal will decay from its initial value $\langle S\rangle_t$ at time t to its final value $\langle S\rangle_{t+\tau}$ at time $t=t+\tau$ with the characteristic time τ of its dynamics, which is

the *relaxation time* by which the signal is expected to become totally uncorrelated. From the simulations of the downhill scenario, autocorrelation of the interstate fluctuations could be calculated to obtain the dynamics of the process. For the simulation shown above, the autocorrelation analysis is shown in Figure 2.6.

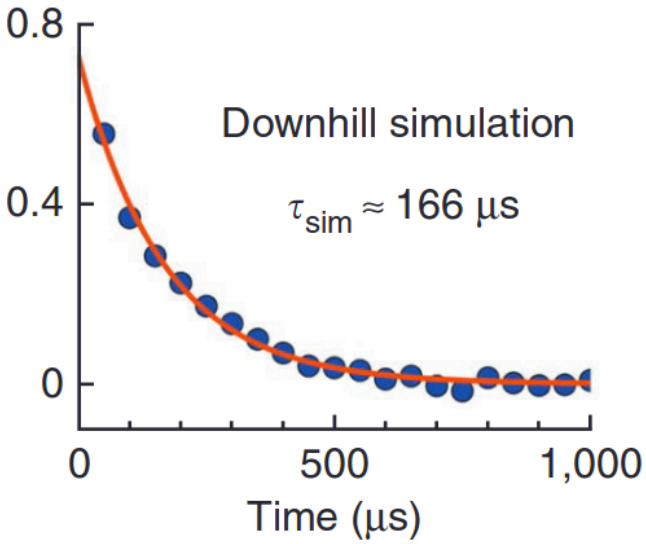


Figure 2.10 Kinetics obtained from the stochastic simulations using autocorrelation of the fluctuations in the modeled signal (FRET efficiency, in this case). Adapted from Campos et al. [74]

2.7 Conclusions

Stochastic kinetic simulations are a faster and simpler approach to study and investigate kinetics of protein folding. Dynamics and mechanistic details of elementary structures in proteins could be studied using the approach. We show that with the stochastic kinetic simulations on α helices where the simulations corroborated observations in the experiments and pointed to the basic motions in helices. Modeling state-to-state transitions stochastically, the dynamics present in the molecule could be unraveled directly from the stochastic realizations in the trajectories. Such trajectories are the kinetically relevant jumps of the molecule and appropriately mapping signals to such states, one could suitably model particular experiments such as FRET, which will be shown in the following chapters. The simulations illustrated here lack structural detail but introducing the stochastic kinetic approach over other simple models containing structural resolution is straightforward. Recent example of employing stochastic kinetic simulations with a simple statistical mechanical model that point to similar folding mechanisms for villin as compared to those from extensive and long all-atom molecular dynamics trajectories [75] illustrate the power and capability of the approach underlined here.

3 Probing the dynamics of single protein molecules with Stochastic Simulations: Single-molecule FRET Studies

In this chapter, the use of stochastic simulations to understand single molecule behavior of proteins and their application in unraveling of the conformational dynamics of single molecules using simple 1-D models of protein folding is described.

3.1 Simple 1-dimensional Free Energy Surface Models of Protein Folding

Based on direct implications of the Energy Landscape Theory, multidimensional free energy landscapes of protein folding could be projected into simpler dimensions using suitable reaction coordinates. Using a mean field approach, Muñoz formulated[49] a simple model referred to as 1-dimensional Free Energy Surface (referred to as 1D-FES or FES hereafter) model that applies this principle to make a 1-dimensional projection of free energy of folding onto a reaction coordinate. It is loosely based on Zwanzig's model[43a] who earlier had developed a simple one dimensional folding model using the number of residues in incorrect conformation as reaction coordinate, even prior to the grounding from the Energy Landscape Theory. The 1D-FES model of folding has been successfully used to analyze and explain different experimental data on protein folding, both thermodynamics and kinetics.

A property termed 'nativeness' n, that is defined as the average probability of finding any residue in native-like conformations is used as the reaction coordinate in the model. 'n' defined this way as a probability is a continuous function, an analogue version of fraction of residues in native conformations.

Conformational entropy is then calculated using simple Gibbs entropy formula:

$$\Delta S_{res}^{conf}(n) = (-R[n\ln(n) + (1-n)\ln(1-n)] + n\Delta S_{res}^{n=1} + (1-n)\Delta S_{res}^{n=0}) \quad (3.1)$$

$$\Delta S^{conf}(n) = N\Delta S_{res}^{conf}(n) \quad (3.2)$$

$$\Delta S_{res}^{conf}(0) = \Delta S_{res}^{n=0}(0) = S_{res}^{n=0} - S_{res}^{n=1} \quad (3.3)$$

$$\Delta S_{res}^{n=1}(1) = 0 \quad (3.4)$$

where at nativeness n = 1 when every residue in the protein is fully folded is taken as the reference state. $\Delta S_{res}^{conf}(0)$ (at n = 0) is the difference in conformational entropy of the residue that is populating all the non-native

conformations and the residue being fixed in a fully native conformation. N is the total number of residues in the protein.

Using a mean-field approach, the enthalpy functional representing the folding stabilization energy is taken as an exponential function of nativeness, n

$$\Delta H(n) = N\Delta H_{res}[(1+(\exp(\kappa_{\Delta H} n)-1)/(1-\exp(\kappa_{\Delta H}))] \qquad (3.5)$$

where ΔH(n) is the average energy of interaction per residue. $\kappa_{\Delta H}$ is a parameter that defines the curvature of the functional. The one dimensional free energy of folding ΔG(n) is simply given by:

$$\Delta G(n) = \Delta H(n) + T\Delta S^{conf}(n) \qquad (3.6)$$

With appropriate scaling terms using the heat capacity $\Delta C_p(n)$ for the temperature dependence of enthalpy and entropy as applied routinely in protein denaturation experiments (see section 2, Chapter 5), the free energy at any temperature ΔG(T,n) is then given by:

$$\Delta G(T,n) = \Delta H(T,n) + T\Delta S^{conf}(T,n) \qquad (3.7)$$

The above relation is then used to calculate the simple free energy profiles at various temperatures corresponding to the experimental conditions. Chemical denaturation of proteins could also be appropriately included into the 1D-FES model, enabling the analysis of experimental data from such experiments as well.

In this simple model, as implied by the Energy Landscape theory, the free energy barrier for folding arises from non-synchronous decay of the loss of conformational entropy and the gain due to the interaction energy. By simply adjusting the exponent $\kappa_{\Delta H}$ of the stabilization energy, the magnitude of the folding barrier could be adjusted.

Following Kramer's like treatment, dynamics (relaxation kinetics) of the system could be captured as diffusion of the protein molecule over the simple projected 1-D free energy profile. The diffusive kinetics in this model is obtained by employing the rate-matrix formalism of Lapidus et al.[27] on a discretized representation of the free energy surface. Rate is primarily determined by D_{eff} the effective intra-molecular diffusion coefficient that is given by:

$$D_{eff} = k_o \exp(-NE_{a,res}/RT) \qquad (3.8)$$

k_o the pre-exponential factor is assumed to be temperature independent. $E_{a,res}$ the activation energy per residue incorporates all the complexities such as viscosity dependence, energy landscape roughness as manifested in the internal friction etc. that are temperature dependent. D_{eff} sets the timescales of dynamics of the diffusion of the molecule on the one dimensional free energy surface.

For proteins with folding barriers larger than ~1kT, rates could also be obtained using an approximation, by calculating the free energy barrier and assuming an appropriate pre-exponential term k_o as:

$$k = k_o \exp(-\Delta G^{\Xi} / RT) \qquad (3.9)$$

where ΔG^{Ξ} is the height of the barrier. As the barrier arising from the asynchronous enthalpy-entropy compensation typically has been found to result in the range of nativeness values of 0.7-0.8[76], the forward and backward barrier heights are calculated by dividing the free energy surface between these values. Value of the pre-exponential k_o scales with size of the protein as $k_o = 3.5x\ 10^6 /N\ s^{-1}$ that is equivalent to the effective diffusion coefficient value of $D_{eff} = 8x10^4x\ n^2/N\ s^{-1}$ in the diffusive kinetic formulation given above. N is the number of residues in the protein. D_{eff} an effective diffusion coefficient (with physical units m^2/s) is given in terms of nativeness with units n^2/s.

This simple phenomenological model has been successively employed multiple times in the analysis of various protein folding experimental data. The systematic deviations from two-state folding behavior in many proteins as observed in chemical and temperature denaturation of proteins was rationalized using the model[77]. Size scaling of stabilities and kinetics of proteins has been analyzed with the model[78]. A large dataset of mutations resulting from ϕ-value analysis of 25 different proteins was analyzed using the model[79]. The model has also been used to analyze a kinetics dataset of 52 proteins to develop an algorithm for predicting both folding and unfolding rates of proteins based only on size and topology to a high accuracy (implemented as a webserver named PREFUR[76]). The model has also been used to analyze quantitatively various biophysical experimental data for the following individual proteins: BBL[80], and its structural homolog PDD[81], gpW[23a] and λ-repressor[82].

The use of this model in analysis of single molecule FRET experimental data forms a major part of this dissertation and a procedure developed for analyzing single molecule time stamped photon measurements is explained in the next chapter. This chapter mainly provides a background to the single molecule experiments and highlights the results of comparing FRET trajectories obtained using stochastic simulations to experimentally measured ones.

3.2 Stochastic kinetic simulations of protein folding:

A protein molecule is modeled as having many discrete microstates with interstate transition rates given by $k_{i->i+1}$ and $k_{i+1->i}$ as forward and reverse (k_f and k_r) is shown in the figure below (Figure 3.1).

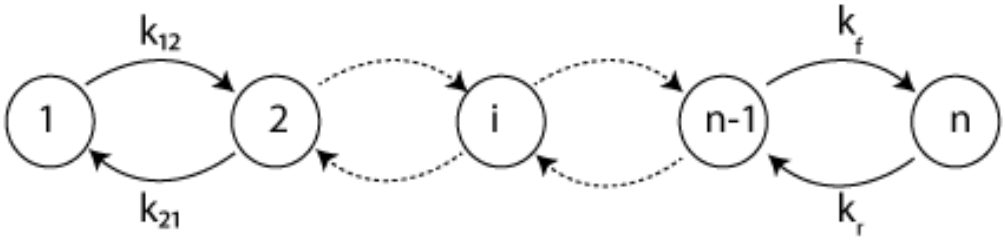


Figure 3.1 Discretized states of a protein with forward and reverse rate constants

The 1-D free energy surfaces obtained from the FES model could be discretized into a number of such microstates. The discretization also provides the probabilities of each of the microstates. As the kinetics is defined as diffusive on the free energy surface, the microscopic rate constants for interstate transitions could be obtained using the rate matrix formalism [27] introduced in the previous chapter for simulating downhill folding. The time dependent probabilities are then given as:

$$p(i \to i+1) = \Delta t \frac{1}{2}\left(\frac{p_{i+1}}{p_i} D + D\right)$$

$$p(i \to i-1) = \Delta t \frac{1}{2}\left(\frac{p_{i-1}}{p_i} D + D\right)$$

$$p(i \to i) = 1 - [p(i \to i+1) + p(i \to i-1)]$$

where D, the diffusion coefficient sets the timescales for the dynamics as described above. p_{i+1},p_i and p_{i-1} are obtained as the discrete state probabilities from the 1D-FES model. For such a system, a small Δt is chosen so as to ensure only one transition at any given time step and such that the probability to jump to any of the flanking states is <0.1. Stochastic realizations of jumps on the free energy surface defined by the model according to the parameters could be obtained from the kinetic simulation using the recipe given in section 2.3. This stochastic kinetic approach could in fact be readily extended to any model where the conformational reaction coordinate q is continuous and could be discretized and the conformational dynamics described by diffusion. The probabilities so obtained could be used along with the appropriate diffusion coefficient to obtain stochastic kinetic trajectories.

3.3 Comparison with smFRET experiments

3.3.1 Background: Single molecule FRET

Förster Resonance Energy Transfer (FRET) is a powerful technique for studying distance distributions and dynamics in biomolecules both intra-molecular and inter-molecular mainly because its sensitivity range of 1-10 nm is very suitable and appropriate for biomolecules. FRET is a non-radiative energy transfer through dipole-dipole interaction occurring between an excited dye referred to as donor and a nearby dye referred to as acceptor, in its ground state (Figure 3.1). In the 1940s Theodor Förster formulated the theory of this non-radiative

energy transfer between suitable dyes in which he showed that the rate of energy transfer is proportional to sixth power of the separation distance between the dyes.

$$k_F = k_D \left(\frac{Ro}{r}\right)^6 \qquad (3.10)$$

where $\tau_D = 1/k_D$ is the excited state lifetime of the donor dye in the absence of the acceptor dye. r is the distance between the two dyes and *Ro* is the Forster radius for the dye pairs, the distance between the dyes with 50% energy transfer efficiency. It is a proportionality constant that depends on the dipole-dipole interactions between donor and acceptor dyes. *Ro* is given by

$$R_o^6 = \frac{9000\,(\ln 10)\,\kappa^2 Q_D J}{128\,\pi^5 n^4 N_A} \qquad (3.11)$$

R_o is typically in the range of 2-9 nm, a very relevant range of the distances of interest in biological molecules, thus making FRET a suitable experimental technique. R_o could be directly obtained from standard spectroscopic measurements without the necessity for any theoretical calculations.

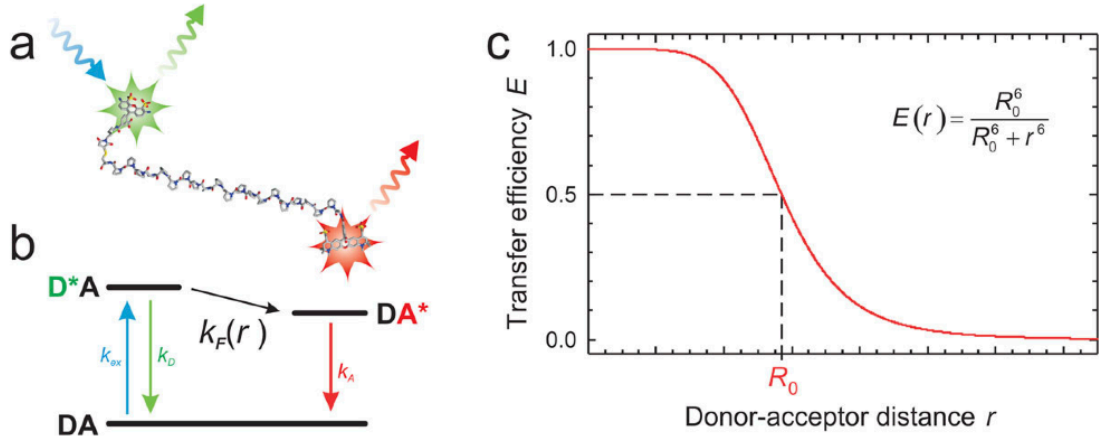


Figure 3.2 Schematic showing the energy transfer process and the typical inter-dye distance vs. transfer efficiency diagram. Dye pairs attached to a protein molecule via linkers undergoing FRET. Adapted from Schuler et al., [83]

Efficiency of the energy transfer depends on the overlap between the donor emission spectrum and the acceptor excitation spectrum that is given by the overlap integral *J*

$$J = \frac{\int F_D(\lambda)\varepsilon(\lambda)\lambda^4 d\lambda}{\int F_D(\lambda) d\lambda} \qquad (3.12)$$

where $F_D(\lambda)$ is the donor fluorescence spectrum and $\varepsilon(\lambda)$ is the acceptor extinction spectrum in molar extinction units ($cm^{-1}M^{-1}$)

Q_D in Eq. 3.11 is the donor fluorescence quantum yield; n is the refractive index of the medium surrounding the dyes and N_A is Avagadro's constant. Orientation factor, kappa is defined as:

$$\kappa^2 = (\cos\theta_T - 3\cos\theta_D \cos\theta_A)^2 \qquad (3.13)$$

where θ_T is the angle between the transition dipoles of donor and acceptor, θ_D and θ_A are the angles between the dipoles and the lines connecting the dipoles to the centers of donor and acceptor dyes, respectively. This is shown in Figure 3.3.

In the cases when the rotational reorientation of the dyes is fast compared to the fluorescence lifetimes of the dyes, κ^2 could be averaged to a value of 2/3. For proteins in aqueous solutions, for typical fluorophores used in FRET studies, using steady state polarization measurement, anisotropies of ~0.05-1 [84] has been confirmed, which validates the assumption of faster reorientation times compared to excited donor lifetimes. Also, this assumption is generally valid given that the dipole rotation times are in the picosecond timescales whereas fluorescence lifetimes are typically in the nanosecond timescales. With the rotational averaging of the dipole orientations, calculating FRET is simplified.

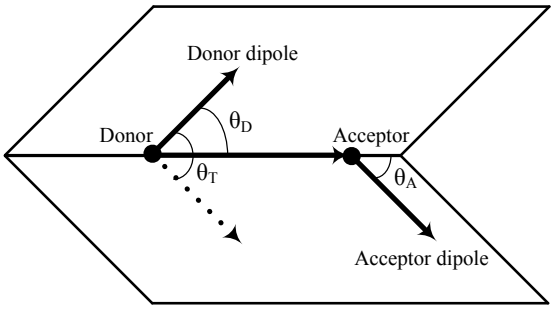


Figure 3.3 Orientation factor κ^2: θ_D and θ_T are the angles between dipoles and the vector joining the donor and acceptor; θ_T is the angle formed between the two dipoles. Typical values of κ^2=2/3 are taken for FRET calculations.

Energy transfer efficiency, the probability that a photon absorbed by donor dye will lead to energy transfer to the acceptor is given by:

$$E = \frac{k_F}{k_D + k_F} = \frac{1}{1 + (r/R_0)^6} \qquad (3.14)$$

Since it is experimentally challenging to directly measure the transfer rate k_F, E is determined experimentally typically by counting the number of emitted donor and acceptor photons using the ratiometric equation:

$$E = \frac{n_A}{n_A + n_D} \qquad (3.15)$$

where n_A, n_D are the counts of acceptor and donor photons respectively. It is also the same as ratio of relative acceptor (I_A) and donor fluorescence intensities (I_D) measured in a time window.

$$E = \frac{I_A}{I_A + I_D} \qquad (3.16)$$

This is the widely used procedure for calculating FRET. Measured n_A , n_D need to be corrected for factors such as differences in quantum yields of the dyes, other photophysical and instrumentation effects such as leak through from donor to acceptor channels, variations in the detection efficiencies etc.

E could also be obtained by measuring the donor lifetimes in the absence (τ_D) and presence of acceptor dye (τ_{DA}).

$$E = 1 - \frac{\tau_{DA}}{\tau_D} \qquad (3.17)$$

This represents the simplest case where there is a single fixed distance between donor and acceptor (no dynamics). It needs to be corrected for if there are distance distributions observed in the molecules in which case the observed deviations could then be used to infer dynamics in the system.

FRET as a "spectroscopic ruler" was first experimentally demonstrated by Stryer and Haugland in 1967 [85] by measuring the distance distributions observed in a bulk solution ("ensemble") of biomolecules and confirmed the Forster relationship of:

$$E = \frac{R_o^6}{R_o^6 + r^6} \qquad (3.18)$$

Nearly 30 years after the first bulk FRET measurements, Ha et al in 1996 used the technique to first measure single molecules of DNA[86]. Since then, (single molecule FRET) has been applied to characterize different biological systems, from protein dynamics, protein-protein interactions to ion-channel mechanisms.

3.4 smFRET Experiments for Protein Folding and Dynamics

For observing single biological molecules of interest using smFRET, foremost requirement is the presence of suitable fluorophores/dyes. Among natural amino acids in proteins tryptophan has the highest intrinsic fluorescence. However it is not very bright having a quantum yield of just 0.13 and also is not very photostable. Hence it is not suitable for single molecule detection. Labeling of proteins with extrinsic fluorophores is therefore the natural choice. Many strategies have been developed for the purpose of labeling proteins with dyes both specifically and nonspecifically. Currently, the most commonly used approach and one of the simplest available is to exploit derivatisation of cysteines using maleimide chemistry [87]. Site-specific cysteines are introduced or removed from the proteins as needed using site-directed mutagenesis and various suitable functionalized fluorescent dyes are attached to the proteins at these locations. Examples of particularly popular dyes for studying proteins are

Alexa Fluor Series [88] owing to their high extinction coefficients, high photostability and high quantum yields. Dyes are typically attached to the protein molecule with linkers such that there is some separation between the fluorescent portion of the dye and the protein chain. For nucleic acids these have been typically cyanine dyes, again because of similarly preferable properties.

Once the protein has been labeled with suitable dyes (acceptor and donor) at specific positions, there are two types of smFRET experiments possible. Simplest type of smFRET experiment to perform is on freely diffusing molecules. A confocal microscope is used for collecting the photons from each of the dyes as a single molecule passes through the 'confocal' volume. The schematic of a confocal instrument is shown in Figure 3.3**.** Using a high numerical aperture objective a laser beam is focused into a tiny spot in the sample solution, the observation volume (confocal volume) of which is typically ~1fl. As the laser wavelength is chosen as appropriate for exciting donor fluorophores, when a free molecule diffuses into this spot, the donor gets excited. It may relax to the ground state by a radiative process (emitting a donor photon) or could transfer its energy via FRET to the acceptor chromophore in the protein. The latter happens when the dyes are within a suitable distance that depends on the dynamics in the protein. In this case, there will be an acceptor photon emitted but no donor photons. The emitted photons are collected using the same objective lens using dichroic mirrors to separate them from the original laser beam used for excitation. Thus only the fluorescent photons are transmitted to the detection system, with another dichroic mirror spectrally separating the donor from the acceptor signal. A confocal pinhole is used to further the spatial selection and thus minimize the unwanted background signal by removing out of focus light. Using a polarizing beam splitter, the collected light can be then split based on polarization for the additional anisotropy measurements. Separated donor and acceptor photons are detected using suitable photo-detectors. Avalanche photodiodes (APD) that are highly sensitive even for single photons are used to obtain counts of the detected photons for each of the channels, using counting electronics with a detection limit of ~50ps time range. The electronics also record the times of photon arrivals and if a pulsed laser is used for excitation of the donor dyes, then the pulse times are also recorded as the reference time points.

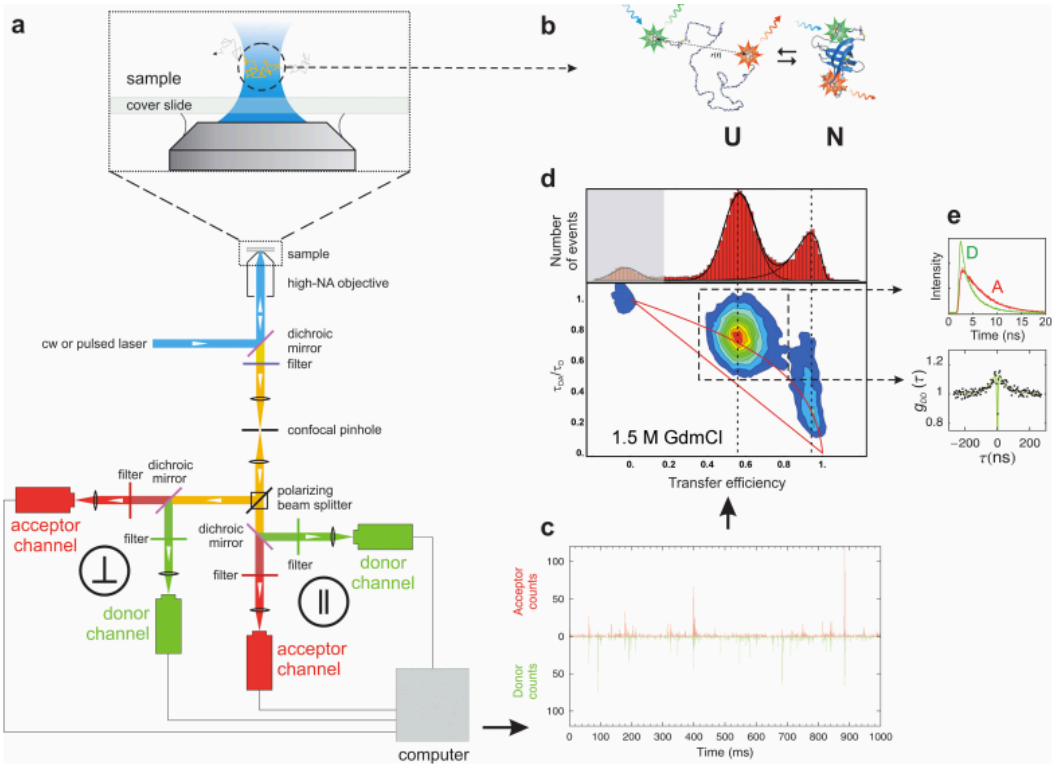


Figure 3.4 Schematic of Confocal Setup and example FRET experiment. a) a 4 channel confocal single molecule instrument that collects fluorescent photons from the dyes separated by wavelength and polarization and records their arrival times schematic b) Protein labeled with acceptor and donor dyes showing folding-unfolding transition. c) Photons recorded from free diffusion experiments by the instrument. A short time bin is shown with donor photons in green and acceptor in red d) FEH and a 2D histogram of lifetimes vs. transfer efficiency e) Single photon counting histograms and donor intensity correlation reporting nanosecond dynamics (Adapted from #Schuler, 2014)

As the concentrations used are very low (10-100 pM ranges) the probability of two molecules simultaneously diffusing into the confocal volume is almost negligible. Ideally, a single labeled protein molecule will diffuse into this volume and depending on its conformational state and dynamics, a sequence of donor or acceptor photons is suitably recorded with time stamps during the time the molecule resides in the confocal volume (diffuses through). This residence time is on average ~0.75-1 ms for proteins under the typical experimental conditions. Typically a few tens to few hundreds of photons are collected within this time window.

The other type of smFRET experiment is where the protein is immobilized onto a surface with a suitable tether that is typically a biotin-streptavidin-biotin linker. This removes the problem of short residence time windows for photon collection when free molecules diffuse into and out of the confocal volumes and thus enables collection of photons for much longer times, until the dyes are photobleached and are no longer fluorescing. Efforts must be taken so as to ensure that the immobilization procedure doesn't affect the conformational dynamics and properties of the protein molecule. Another variant of the immobilization procedure is to encapsulate single protein molecules into vesicles and attaching these vesicles to the surface. [89]

With the multiparameter fluorescence detection setups that are typical nowadays, the data recorded from an experiment contain the information on color, polarization and arrival time of each of the individual photons resulting in photon trajectories. The arrival times are both absolute and relative to the laser pulse used for exciting when using a pulsed laser setup. From such information, different quantities of interest such as fluorescence lifetimes of the dyes, mobility of the dyes (from fluorescence anisotropy), and FRET efficiencies from intensity ratios, etc. could be derived.

A straightforward analysis is plotting the counts of donor and acceptor photons detected within a suitable time window or bin e.g. 1 ms. This typically shows long time bins of a base low level signal (background photons) and sudden increases in the detection counts referred to as 'bursts' of photons. These bursts are detected simply by applying a threshold number of photons in the time bins although a number of other sophisticated methods for bursts detection have been developed [90]. A typical smFRET experiment is carried out for hours, collecting thousands of such bursts. A simple way to analyze these bursts (or bins in the case when a simple binning and threshold procedure is used) is then to calculate FRET efficiency E ratiometrically from the photon counts in the bins (as given by equation 3.15). The distribution of FRET efficiency from all the bursts is called as FRET Efficiency Histogram (FEH). From the number of peaks and the position of the peaks in FEH, the heterogeneity in the states of the molecules corresponding to the peaks could be inferred. A peak is typically observed at transfer efficiency values of ~0 which is referred to as zero peak and results from donor-only molecules passing through the confocal volume. Such molecules could result from imperfect labeling of the protein or photobleaching of acceptor dyes etc. and therefore lack suitable acceptor dyes to receive the transfer energy from donor thus only emit donor photons. Example FEH for a protein is also shown in Figure 5.4. For a labeled protein undergoing conformational fluctuations between the unfolded and folded states, FEH will show two distinct peaks with the positions corresponding to FRET efficiencies of each of the states. Theories of FEH have been developed and even analytical formulations are available for simple systems containing two-states (two peaks in the FEH)[91].

3.5 Photon Statistics and Broadness of FEH

Photon emission is a stochastic process and the mean emission rates (or the corresponding mean count rates, to an approximation) are given by a Poissonian distribution. Such discrete stochastic processes give rise to fluctuations in the photon count rates around their mean values for single molecules resulting in a distribution of the resulting FEH that is called as "shot noise". That is, even if the inter-dye distances are fixed single values, there will be distribution observed for the FRET efficiency due to the intrinsic stochastic nature of the photon emissions. The variance in the efficiency distribution due to shot noise is given by:

$$\sigma_{sn}^2 = \langle E^2 \rangle - \langle E \rangle^2 = \langle E \rangle (1 - \langle E \rangle) \left\langle \frac{1}{N} \right\rangle \leq \langle E \rangle (1 - \langle E \rangle) \Big/ N_T \qquad (3.19)$$

where $\langle E \rangle$ is the underlying true mean transfer efficiency, $\left\langle \frac{1}{N} \right\rangle$ is the mean reciprocal of number of photons in the burst and N_T is the minimum burst size which is same as the threshold applied in the burst selection. The implication is that a broad efficiency distribution doesn't necessarily conclude broadness in the distance distribution. This broadening of FEH due to shotnoise is a well-understood phenomenon.

However, it is common to observe FRET efficiency distributions that are broader (Eq. 3.20) than expected just from the contributions of shotnoise alone. The origin of this excess broadness is not understood in full.

$$\sigma^2_{obs} = \sigma^2_{sn} + \sigma^2_{non-sn} \qquad (3.20)$$

One obvious factor is the existence of distance distributions that is typically observed in the broader efficiency distributions for unfolded states that have a large conformational heterogeneity compared to the folded state. Factors such as sample heterogeneity due to labeling differences or other photophysical effects could also contribute to the broadening and needs to be considered. Another most important factor contributing to this excess broadness in the distributions is the conformational dynamics of the protein molecule itself. If the timescale of such conformational dynamics in the protein is similar to the binning times used for the efficiency distributions, excess broadness results. If this is the case, the broadness should depend upon the bin times used and should decrease with the increase of bin times. This procedure has been used to confirm the presence of fast conformational dynamics in some proteins [92]. But, it has severe shortcomings as the method itself is limited by the number of photons that could be collected within the short timescale of the fast process of interest to be studied. For example, if the process has dynamics in 100μs timescale, the effects of binning the already low number of photons that could be collected within this short time of 100μs limits the possible resolution of the process. For faster conformational processes happening in the sub-millisecond regimes, obtaining sufficient number of photons so as to be able to resolve them is typically very difficult. In the following chapter, we develop a methodology to overcome the necessity for binning and directly obtain relevant properties from photon arrival data itself.

3.6 Timescales, FRET Distributions and Dynamics

Photon statistics and resulting FRET efficiency distributions are influenced by many different stochastic processes that occur in a range of timescales. Mainly, these include:

a) photophysical processes such as excitations, radiative and non-radiative decay, energy transfer and photoblinking
b) orientation dynamics of the dyes
c) translational diffusion of the molecule through the laser spot especially if the laser beam has a nonhomogenous illumination profile

d) conformational dynamics of the molecule

These could be classified as relatively fast or slow processes depending on how they compare with the measured interphoton times which are typically in the ~μs timescales. Photophysical processes and the dye reorientations (covered in section 2.3) are typically fast processes as they happen in the ps-ns range.

The effect of translational diffusion on experimentally observed FRET efficiencies has been taken into account rigorously [93]. As covered in the previous section, for FRET efficiency distributions produced by analyzing photon trajectories with bins of equal durations, processes that are slower than the interphoton times manifests as the extra width or broadening over the shotnoise in the distributions. For burst analysis, when conformational changes or dynamics do not occur during the time the molecule traverses the laser spot the effect of diffusion could be bypassed in the analysis and need not be modeled explicitly as the total countrates are not affected. It has been shown [93] that for a single molecule observed in the laser spot, the analysis of free diffusion FRET experiments could be simplified if: FRET efficiency is independent of the location of the molecule in the spot and the total photon count rate (sum of donor and acceptor photon count rates) is same for all the conformations. Practically these conditions hold in typical FRET experiments thus obviating the explicit modeling of effects of translational diffusion in FEH analysis.

Conformational dynamics of the molecule is the process of interest to be studied from the FRET experiments. Of these, several processes like linker fluctuations, fast dynamics such as vibrations etc. occur in sub-microsecond timescales. Only the processes occurring in timescales similar to or slower than the observation time (interphoton times) (~μs) affecting the inter-dye distances and thus the count rates and transfer efficiencies are relevant. From the observed photon trajectories and FEH these conformational dynamics are what is of interest to be extracted and determined.

3.7 Binning Times and Effects on Folding Scenarios

For different folding scenarios and timescales of dynamics, the effect of binning times used in the construction of FEH is different. The differences critically depend on the interplay between the binning times T_b and the relaxation time τ of the dynamics in the protein molecule. This is shown in Figure 3.5 for three different folding conditions.

For a two-state folding scenario, the protein always crosses a free energy barrier to populate the other state from its current state that could be native or unfolded. Hence there are always two peaks that must be observed in the FEH with their amplitudes determined by the equilibrium constant when the process could be totally resolved i.e. $T_b << \tau$. The widths of both the peaks should be shotnoise limited and there should not be any broadening observed. However, when $T_b >> \tau$, there is dynamic averaging due to multiple folding-unfolding

transitions within the observation time (T_b) and thus a single averaged peak is observed. For the one-state scenario, there is always only one peak observed in

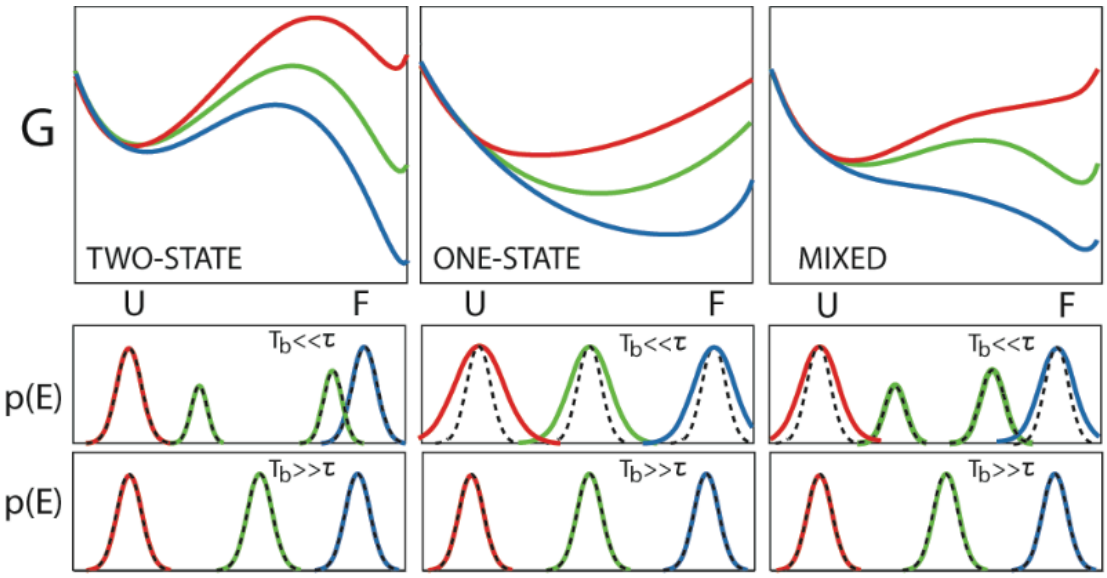


Figure 3.5 Three different folding scenarios and the effect of binning times on observed FEH. Native, midpoint and unfolding conditions are shown in blue, green and red, respectively. G is the free energy and p(E) is the FRET efficiency. Profiles simulated under different conditions using 1-D FES model.

the FEH. When $T_b << \tau$, the peak shows more broadening than the shot noise defined width whereas when $T_b >> \tau$, the width of the peak is shotnoise limited and thus indistinguishable from the two-state case. In the marginal barrier scenario with significant barriers only under midpoint conditions, there are single peaks broader than shotnoise defined widths in both folding and unfolding conditions and two peaks that are shot noise limited at the midpoint condition when $T_b << \tau$. Thus it is a mixed behavior depending on the conditions. This highlights the criticality of having enough photons to reliably construct FEH and use Tb that are much lesser than τ. For fast folding proteins with already small τ, this is not possible and pushes the limits of the smFRET technique and also warrants better methods of analysis than FEH. Such a procedure that increases the time resolution of the technique is presented in the following chapter.

3.8 Stochastic simulations and Single molecule behavior

Using the simple free energy surface model described in section 3.1, different scenarios of folding, as a two-state process, a marginal barrier scenario and a one-state scenario were modeled with three different barriers (~4RT, 1RT and 0RT, respectively) by varying the parameters. Stochastic simulations were performed for observing the behavior of the molecules under these scenarios.

The simulations reveal directly the behaviors that could be observed in single molecule measurements for proteins folding under such different scenarios. For the two-state case, the trajectories clearly point that the molecule will be fluctuating preferentially between the discrete microstates that correspond to folded state and those corresponding to unfolded states. As a sharp boundary separating the two macro-states (folded and unfolded wells) could be drawn to

distinguish them, the stochastic trajectories directly and unequivocally point that the system has two distinctive states.

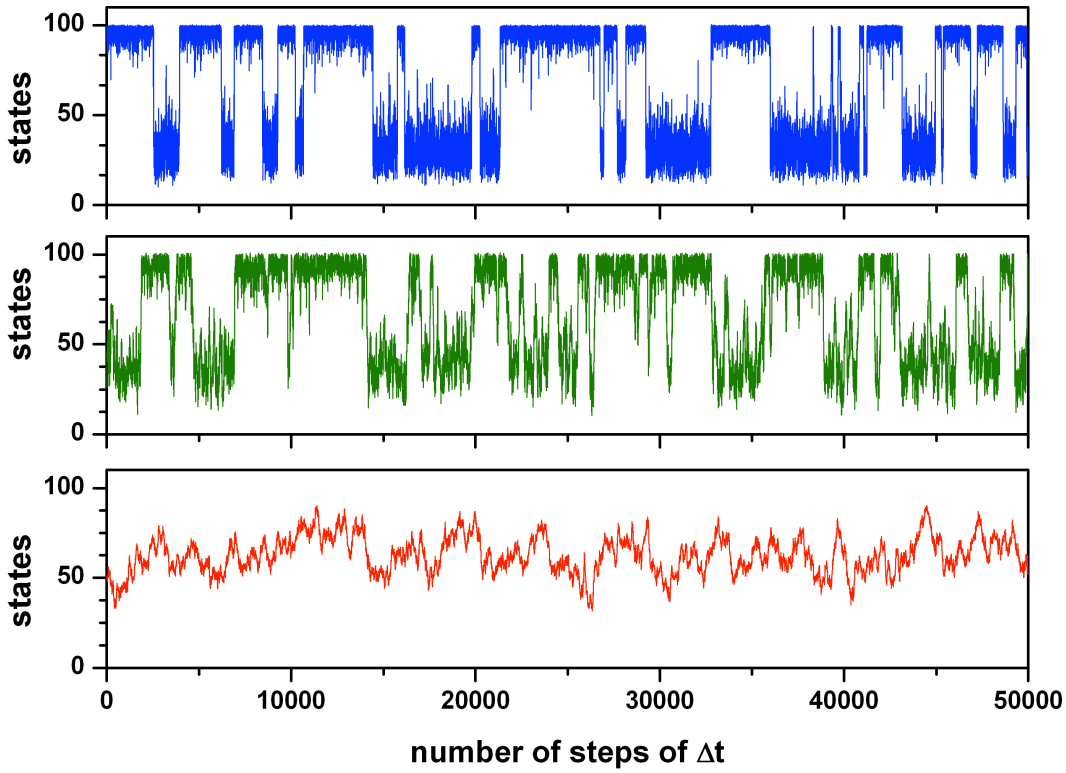


Figure 3.6 Stochastic trajectories for three different folding scenarios. Blue, green and red show simulations of two-state, marginal and downhill scenarios.

In the marginal barrier scenario, the trajectories reveal a very interesting behavior. Even with a small barrier of only ~1RT the molecule has a majority residence time populating the distinctive macro states (folded and unfolded wells) if one defines a sharp boundary of separation. However, the trajectories clearly show that the molecule populates the intermediary microstates often. These could be referred to as 'excursions' of the molecule on the free energy well. The trajectories point out that in such a scenario 'intermediates' could be observed in the single molecule measurements even though from bulk experiments it may not be the case. Such single molecule hopping behavior over small entropic barriers have been observed when using force to unfold a protein. Under constant force atomic force experiments, for gpW, a protein with a very small barrier such hopping has been observed as shown in Figure 3.7. (Schönfelder et al. unpublished results)

For barrierless one-state folding scenario, the trajectory is reflective of the Brownian motion like diffusive nature of the process. Without a barrier, the protein simply fluctuates around the energy surface populating microstates that are energetically favored under the given conditions. These could also be described as 'excursions' on the free energy surface. This behavior is similar to the fluctuations of a spring on a simple harmonic potential as shown in the previous chapter (Figure 3.8)

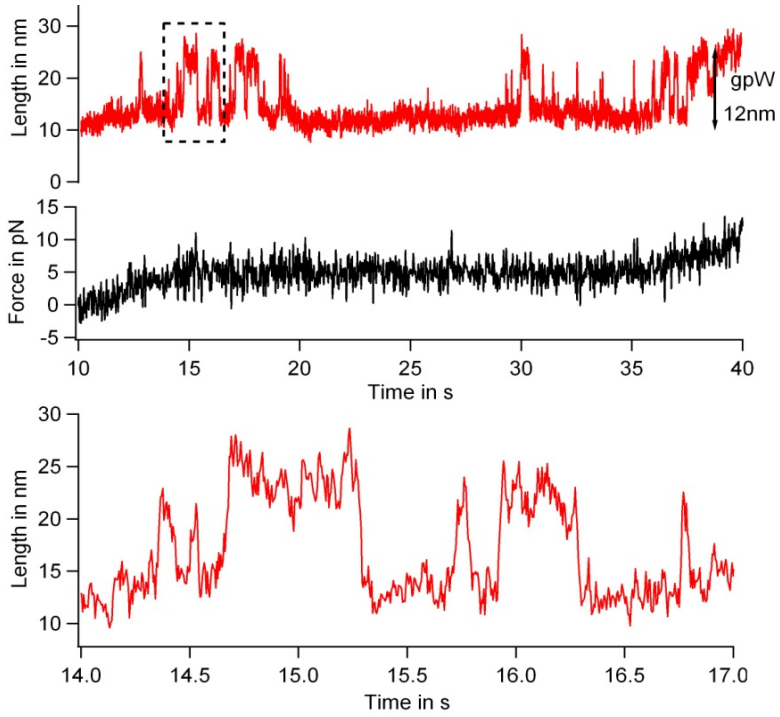


Figure 3.7 Hopping behavior revealed in the stochastic simulation trajectories for marginal barrier scenario are observed in constant force measurements of gpW (5pN, time extended), a marginal barrier protein. Comparisons with smFRET experiments [Schönfelder et al. unpublished data]

Stochastic kinetic simulations could be used to model and study experiments by modeling and simulating appropriately the experimental signals that are to be analyzed. Here, we simulate FRET efficiency by using Poissonian random variables to simulate photon emissions depending on the microstate the molecule is in at each time step (Δt). The photons are emitted at a set photon flux which determines the average number of photons emitted in a given time window. From the counts of donor and acceptor photons emitted FRET trajectories could be generated for direct comparisons with the experimentally observed ones. The photon flux is then finely adjusted to reproduce the same average of the reciprocal of the total number of photon observed experimentally within short time bins and trajectories of lengths comparable to those from experiments were generated. Here, we show such a simulated trajectory of 10ms, closely matching the experimentally observed ones for BBL, a one-state folding protein. The experimental trajectories were obtained from free diffusion experiments under midpoint conditions of 5M urea with high photon count rate. The simulated trajectories (shown in blue) were generated with a flux of 0.8MHz of photon emissions and they overlap well with the experimental trajectories shown as cyan circles. The FRET fluctuations in the stochastic trajectories are similar to the ones observed in the experimental ones in both timescales and amplitudes. The good match between stochastic simulation trajectories and the experimental observations is highlighted further by the apparent incapability of even a seven state model to reproduce such fluctuations of downhill folding BBL as shown by the red line that is a fit to this model.

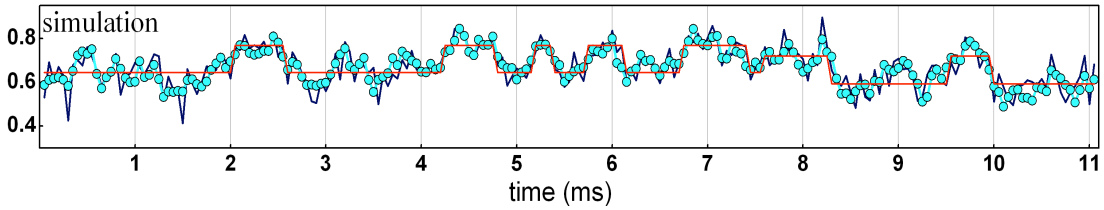


Figure 3.8 Comparison between simulated and experimental FRET trajectories for BBL obtained from free diffusing experiments.

3.9 Conclusions

Here we have shown that the simple stochastic simulations could reveal the single molecule behaviors of proteins folding under different scenarios. The effects of barrier, even if small (~1RT) are pronounced as the molecule displays 'excursions' on the free energy surface as shown in the trajectories for the marginal barrier case. For downhill scenario, the trajectories reveal Brownian motion like diffusive motions on the surface. These simulations offer a fast and effective method to make comparisons with experimental observations, which is demonstrated by comparisons with FRET data.

4 Decoding conformational dynamics and free energy surfaces of fast-folding proteins from single molecule Photon Arrival Trajectories

Chapter adapted from the article titled "A procedure for deriving the folding free energy surface and conformational dynamics of fast-folding proteins from single molecule photon arrival trajectories" under preparation.

4.1 Introduction

Advances in single molecule experimental techniques have led to an increasing number of their applications in the study and characterization of bio-molecules. The last 15 years have marked the beginning of a new era in molecular biophysics based on the explosive growth of single-molecule techniques to study biomolecular processes.[94] Among these, single molecule Förster Resonance Energy Transfer spectroscopy (smFRET) is of particular interest because it is directly comparable to standard bulk experiments thus allowing for mutual cross-checks. In the area of protein folding, smFRET has the potential to reveal fundamental single molecule properties such as population distributions and conformational dynamics. [83, 95] In the last years, smFRET has been successfully applied to a variety of problems related to protein folding, including the demonstration of two-state folding [96], the dimensional analysis of unfolded states [97] and intrinsically disordered proteins[98], the single molecule characterization of one-state downhill folding[22], and very recently the study of folding transition path times[33]. The accumulation of experimental studies in this area of biophysics has been accompanied by the parallel development of methods for analyzing and interpreting these new data quantitatively.[99]

The most typical smFRET experiment to study protein folding has been to label the protein with appropriate donor and acceptor dyes and collect photons emitted by these fluorophores while the molecule is freely diffusing under the illumination volume of a confocal microscope (two-colored FRET on free diffusing molecules)[95]. Another kind of smFRET experiments involves immobilizing the protein to a surface, which then allows collecting the emitted photons (two colored FRET on immobilized molecules) for times much longer than the observation times in the free diffusion experiments (typically <1ms) [95]. The output measurements of such experiments are termed **photon trajectories** and correspond to sequences of photons for which the color and arrival time to the detector are recorded with picosecond resolution. Individual detected photons, however, are typically interspersed in intervals of several microseconds

due to inherent limitations in instrument detection efficiency as well as in the excitation and emission rates of organic fluorophores.[94a] This data is typically binned in intervals ranging from 0.1 to 1 milliseconds to produce histograms of averaged photon counts (Photon Counting Histograms), which are then converted into FRET efficiency histograms (FEH). In these cases FRET efficiency is simply defined as the ratio of acceptor photon counts to the total counts in each bin. The overall FEH shape and the mean values of the observed peaks provide information on the sub-populations, their structural properties from the inter-dye distance distributions within them, and the interchange kinetics between the subpopulations from their dwell time distributions[92]. (This has been described in detail in the previous chapter).

Photon emissions being fundamentally stochastic there is always a width in the FEH due to the limited photon statistics within a given time window that is referred to as shot noise. Shot noise contributes to noisier measurements in shorter bursts of photons collected from free diffusion experiments than longer ones, thus making it preferable to collect more photons and for longer durations. The effect from shot noise is exactly quantifiable but when there is excess width observed in the FEH more than accountable by shotnoise alone, it is hard to characterize the exact causative factors and further analysis is warranted. Binning of the data also poses a limitation in the FEH analysis. When the timescales of conformational dynamics is similar to or faster than the bin times, it is no longer possible to clearly distinguish the different states of the molecule due to the dynamic averaging similar to what is called line broadening in NMR experiments. Analyzing the shape of the FEH and its dependence on the bin times is then needed to obtain the kinetics information from the data.

Much progress has been made in the analysis of the FEH with various methods [92] [100] [101] [102] and even analytical formulations [103] [104] [91] albeit only for simple two-state models have been developed, but when the conformational transitions are as fast as the burst/bin durations especially for free-diffusion experiments the procedures involve many approximations and key assumptions such as the independence of countrates on the translational diffusion of the molecule across the laser spot [105].

With the discovery and characterization of many fast folding proteins that have relaxation times less than <1ms [25] and the identification of ultrafast one-state folders [20], the conformational dynamics and folding motions of the protein molecule that needs to be studied are in μs range. For characterization of their fast dynamic processes the time resolution of smFRET technique has to be drastically improved [106]. The time resolution of smFRET is primarily limited by the number of photons emitted by the fluorophores and their efficient detection, each of which has practical limitations. The dyes suffer from various photochemical and photophysical problems such as blinking, photobleaching etc. that limit the photon fluxes obtainable. With advances such as chemical cocktails for photo-protection from bleaching[74] and other methods[107] to counter these problems and with the improvements in the instrumentation for increasing detection efficiencies [108], the accessible time resolution has been greatly increased. Photon fluxes of ~1 photon per μs, a hard empirical limit for the

currently available dyes and detectors, when successfully reached enable the time-resolved study of many fast folding proteins and other processes. In the few experimental studies in which the photon fluxes nearer to this limit were reached[22], the typical analysis methods based on FEH had the broadening problems mentioned above since the conformational dynamics of the studied proteins was faster than the bin times that could be applied.

In such cases, methods that directly utilize all the information available in the time stamped photon trajectories for extracting the kinetics and dynamics become necessary. Many such methods have also been developed and successfully applied for characterizing various biological processes [109] [110] [111] [112] [113] [114] [115] [116]. Most of these methods are based on a maximum likelihood approach while some of the Hidden Markov Modeling (HMM) based methods [115] [117] [118] still involve binning or converting the photon trajectories into FRET efficiency trajectories (instead of histograms) and thus suffer from the inherent drawback of limited time resolution or with inaccuracies resulting from converting short photon trajectories into FRET efficiency trajectories. Also, the effect of fluctuations of photon emission rates when the molecule is freely diffusing in the laser spot makes it challenging to decipher its conformational fluctuations occurring at the same time scales from short trajectories obtained from free diffusion experiments[119]. Methods that involve photon count rates typically make a key assumption that the photon emission rates are independent of the translational diffusion of the molecule [105].

Gopich and Szabo have developed a rigorous maximum likelihood based method for analysis of photon trajectories that does not require any binning and doesn't involve photon count rates thus making it equally applicable for trajectories obtained from both type of two-color smFRET experiments (free diffusion and immobilization) [120]. The method involves analyzing the trajectory photon by photon and calculates the likelihood of the observed trajectory being explained by a given kinetic model with its parameters. It can be used for both kinetic model selection and for identifying the parameters for a given model. The method is model dependent and for simpler models such as two-state model the likelihood calculation is indeed exact. The method has been applied for extracting folding and unfolding rate coefficients from experimental single molecule measurements for different protein molecules such as protein α_3D [112], villin subdomain[121], for estimating the upper bounds of transition path times [122] and has been instrumental in the breakthrough experimental characterization of transition path times recently[33]. This rigorous method has so far been only applied with chemical kinetic models such as two or three state models, though it could be extended to diffusive models that capture dynamics of the systems as diffusion on simple 1-dimensional coordinates [106, 120].

The projection of multi-dimensional energy landscapes of protein folding into single dimension using appropriate reaction coordinates and modeling folding as diffusion on this 1-D surface has been instrumental in the study of protein folding. Simple statistical mechanical models with solid theoretical grounding on Energy landscape theory[16] offer a powerful, rigorous and very tractable alternative to the chemistry based models traditionally used in protein folding.

Simple Free Energy Surface Model is one such model that has been successfully applied numerous times to study different aspects of protein folding such as stabilities [51] and kinetics [123], analysis of DSC experiments, predicting of folding and unfolding kinetics[76] as well in the analysis of single molecule experimental data [22].

In this work, we combine the maximum likelihood method with a discretized version of the continuous simple one dimensional model to develop a procedure that directly estimates protein folding thermodynamic parameters (ΔH and folding barriers $\Delta G^{\ddagger}$) and the dynamic parameter which is the intramolecular diffusion coefficient from smFRET photon arrival trajectories. We perform stochastic kinetic simulations on the discretized version of the simple free energy surface model to generate conformational state transition trajectories and photon emissions. We test the procedure thoroughly on synthetically generated trajectories with fully known dynamics and the thermodynamic parameters to study its performance. Effects of the amount of data required to reliably extract the parameters, effect of variations in number of photons arriving within the relaxation times of the process and the effects of the quality of available data with varying amounts of background noise are studied.

4.2 Methods

In this section, we describe the procedure used for combining simple, continuous 1-D Free Energy Surface models of protein folding with the maximum likelihood method for analyzing single molecule photon arrival trajectories. The stochastic kinetic approach used in these simulations to generate data for the evaluation and extensive testing of the new combined procedure is also described.

4.2.1 Combining Simple Free Energy Surface model with the maximum likelihood method

The simple free energy surface model calculates 1-D free energy profile of a protein by using a phenomenological local order parameter called 'nativeness' as reaction coordinate. This order parameter is continuous and represents the 'foldedness' of a protein molecule corresponding to the probability of a number of peptide bonds to be in the native dihedrals. The model has terms for entropic and enthalpic contributions that scale linearly with the number of residues (N) in the protein and are defined as functions of the order parameter as:

$$\Delta S^{conf}(n) = N(-R[n\ln(n)+(1-n)\ln(1-n)]+n\Delta S_{res}^{n=1}+(1-n)\Delta S_{res}^{n=0}) \qquad (4.1)$$

$$\Delta H^{total}(n) = N\Delta H_{res}[(1-x^{(1-n)})/(1-x)] =$$

$$N\Delta H_{local,res}[(1-x_{local,res}{}^{(1-n)})/(1-x_{local,res})]+N\Delta H_{nonlocal,res}[(1-x_{nonlocal,res}{}^{(1-n)})/(1-x_{nonlocal,res})] \qquad (4.2)$$

$$\Delta G(n) = \Delta H^{total}(n) + T\Delta S^{conf}(n) \qquad (4.3)$$

ΔS_{res}^{conf} the entropy cost of fixing a residue in its native conformation is set to a constant value of 16.5 J per mole per K. We use the version of the model using a Markov-chain based formulation [76] where the energy term is further divided into local $\Delta H_{local,res}$ and non-local $\Delta H_{nonlocal,res}$ contributions per residue. x is a characteristic rate of breaking stabilizing native contacts in a Markov chain and the fraction $[(1-x^{(1-n)})/(1-x)]$ gives the remaining stabilization energy as folding progresses. The balance between the enthalpy and entropy functionals result in the free energy (ΔG(n)) whereas the local and nonlocal enthalpy terms determine the magnitude of folding barriers in the model as the specific curvatures of these functionals are kept constant by fixing $x_{local,res}$=3.490 and $x_{nonlocal,res}$=0.002 respectively. These values for fixing the curvatures of the energy contributions were obtained previously from fitting the kinetics of an experimental dataset of 52 proteins using this model. This version of the model with fixed curvatures for the energy terms was chosen, as it is easier to modulate the barrier heights by just changing two parameters. The magnitudes of $\Delta H_{local,res}$ and $\Delta H_{nonlocal,res}$ vary from protein to protein and a range of their values encompass all possible scenarios of varying barrier heights, from global downhill to activated processes with high barriers. Different folding scenarios such as two-state, marginal barrier or global downhill folding are modeled by choosing appropriate values for these parameters (as given in Table 4.1).

The kinetics of the process is modeled as diffusion of the molecule on the 1-D free energy surface and is described by a rate matrix according to Szabo's formalism [27]. For a 100-point discrete model (the order parameter nativeness divided into a 100-point space) the time evolution of the system is given by the rate matrix K

$$\begin{pmatrix}
-\frac{1}{2}D\left(\frac{p_2}{p_1}+1\right) & \frac{Dp_2}{2(p_3+1)} & 0 & 0 & .. & .. & .. & .. & 0 & 0 \\
\frac{1}{2}D\left(\frac{p_2}{p_1}+1\right) & -\frac{1}{2}D\left(\frac{p_2}{(p_3+1)}+\left(\frac{p_3}{p_2}+1\right)\right) & . & 0 & 0 & . & . & . & 0 & 0 \\
0 & \frac{1}{2}D\left(\frac{p_3}{p_2}+1\right) & . & . & 0 & . & . & . & 0 & 0 \\
0 & 0 & . & . & 0 & . & . & . & 0 & 0 \\
0 & 0 & 0 & . & .. & . & . & . & . & 0 \\
0 & 0 & 0 & 0 & 0 & .. & . & . & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & . & .. & . & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & . & . & .. & \frac{Dp_{n-1}}{2(p_n+1)} & 0 \\
0 & 0 & 0 & 0 & 0 & . & . & . & -\frac{1}{2}D\left(\frac{p_{n-2}}{(p_{n-1}+1)}+\left(\frac{p_n}{p_{n-1}}+1\right)\right) & \frac{Dp_{n-1}}{2(p_n+1)} \\
0 & 0 & 0 & 0 & 0 & . & . & 0 & \frac{1}{2}D\left(\frac{p_n}{p_{n-1}}+1\right) & -\frac{1}{2}D\frac{p_{n-1}}{(p_n+1)}
\end{pmatrix}$$

where D is the intramolecular diffusion coefficient that totally determines the timescale of the dynamics and p_i is the probability of state i (each discrete point in nativeness is taken to be a state). This rate matrix can then be diagonalized to obtain the eigen spectrum from which the relaxation rate is directly obtained. An appropriate value of the diffusion coefficient D is chosen for each of the folding scenarios (as given in Table 4.1) to produce similar timescales of dynamics ($\tau \approx 200$ μs) for sake of comparison between them.

4.2.2 Stochastic simulations of conformational transitions and photon emissions

We perform stochastic simulations of the conformational dynamics of the molecule on the free energy surface with photon emissions to generate trajectories of transitions between the states and of photon emissions. The algorithm is similar to that of Gillespie's (described in Chapter 2) and is given as follows. For a given time step Δt, there are only four possible events namely:

- emission of acceptor photon (with count rates n_{Ai})
- emission of donor photon (with count rates n_{Di})
- transition from i to i+1 ($k_{i->i+1}$)
- transition from i to i-1 ($k_{i->i-1}$)

For the molecule in state i, the photon emission rates are Poissonian and are specified by the count rates (nAi and nDi). Successive events are generated by drawing time steps (Δt) from an exponential distribution $\exp(1/k_T)$ and the events are randomly picked according to the probabilities given by $\left[n_{Ai}, n_{Di}, k_{i->i+1}, k_{i->i-1}\right] \times 1/k_T$ where $k_T = n_{Ai} + n_{Di} + k_{i->i+1} + k_{i->i-1}$. The elementary rate constants for the transitions are taken from the rate matrix and the initial state is chosen randomly according to the equilibrium probability **p0**. The acceptor and donor countrates for each state **i** are obtained by multiplying the total countrate (which is set to 800 photons per milliseconds) with its FRET efficiency εi calculated from a linear mapping of inter-dye distances to the *nativeness.* This linear mapping has three parameters: *Ro* which is the known Forster radius for the given FRET dye-pairs, distances *dfu* and *du* which are the slope and intercept of the distance dependence on the *nativeness*. The distance *du* could be taken as the end-to-end distance between the dye pairs when the protein is in the unfolded state and *dfu* is the slope of change in inter-dye distance as the *nativeness* increases to 1. These parameters are set *a priori* based on the experimental conditions generating a given photon trajectory data such as the FRET pairs used in the experiment and size of the studied protein. In the simulations, values of these parameters were set to 5, 3.5 and 3.5 respectively. The output from the simulations are the state trajectories containing the fluctuations of the molecule on the free energy surface and the strips of photons emitted stochastically depending on the count rates (photon trajectories). By varying the number of steps, we can obtain state and photon trajectories of different lengths. We generate long trajectories for yielding the desired number

of photons at specified count rates and simulate bursts of photons in free diffusion experiments by taking many short pieces of 1000 photons each. This provides us longer sampling of the conformational dynamics in the corresponding state trajectories. For example, to obtain a total of 100,000 photons at 800 ms^{-1} we perform a 125 ms long stochastic simulation and divide the resulting photon trajectory into 100 pieces of 1000 photons each that we refer as bursts.

4.2.3 Maximum likelihood method for identifying model parameters from photon trajectories

For a photon burst from a freely diffusing molecule (or a trajectory from an immobilized molecule) the likelihood that the conformational dynamics as measured in the inter-dye distances of the underlying distinct states is described by the simple 1-D Free Energy Surface model is given by:

$$L_t = 1^T \prod_{j=2}^{N} [F(c_j)\exp(K\tau_j)]F(c_1)p_{eq} \qquad (4.4)$$

$F(acceptor) = \mathrm{E}$ and
$F(donor) = \mathrm{I} - \mathrm{E}$

where I is the identity matrix and E is a diagonal matrix of FRET efficiencies of the microstates. p_{eq} is a vector of equilibrium probabilities. τ_j is the inter-photon arrival time between j-1 and j^{th} photon. is the color of the first photon in the trajectory and c_j is the color (indicating whether it is an acceptor or donor) of the j^{th} photon. After N successive matrix-vector multiplications, a final multiplication by 1^T sums the product over all conformational states to yield the likelihood L_t. For multiple bursts (different photon trajectories) the total likelihood is calculated as $\ln L = \sum_t \ln L_t$ by summing the log of individual likelihoods for each of the bursts (trajectories). By maximizing the total likelihood, the most likely parameters of the underlying model are identified. In our implementation, we have utilized eigenvalue formulation by diagonalizing the rate matrix K for speeding up the computations as explained in Gopich,2009 [120]. Since the shape of the 1-D FES is specified by the magnitudes of the energy parameters and the conformational dynamics is specified by the intramolecular diffusion coefficient D, by calculating the total likelihood for various combinations of these three parameters a region of local minima could be identified. We first perform an extensive grid search in the parameter space for identifying local minima by varying the FES model parameters. We then do a bound parameter optimization within the region of the local minima using a simplex algorithm (as implemented in the *fminsearch* algorithm in Matlab (Mathworks Inc, USA) to identify the most likely parameters of the free energy surface model.

4.3 Results and Discussion

We model three different folding scenarios namely two-state, marginal barrier and one state downhill with the choice of parameters as shown in Table 1. The 1-D free energy profiles and the corresponding probabilities are shown in Figure 4.1. These scenarios vary mainly in the height of the folding barriers as the dynamics are calibrated to be on similar timescales of ~200μs with appropriate scaling of the diffusion coefficient D, for protein with 50 residues.

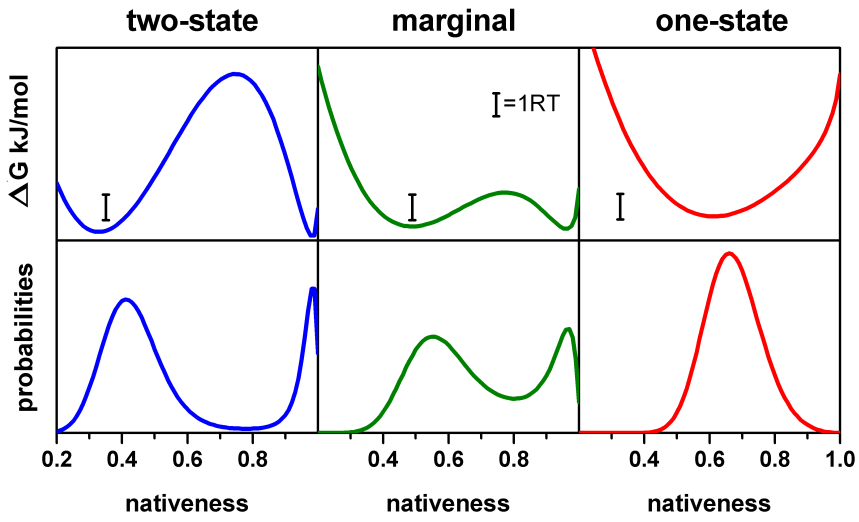


Figure 4.1 Simple free energy profiles and corresponding Probabilities for three different folding scenarios.

For each of the scenarios, stochastic simulations are performed as described in the Methods to generate conformational state trajectories and photon trajectories. Distributions derived from 20s total simulations, sample stochastic state trajectories of 25 ms and 100 μs photon trajectories are shown for each of the scenarios in Figure 4.2.

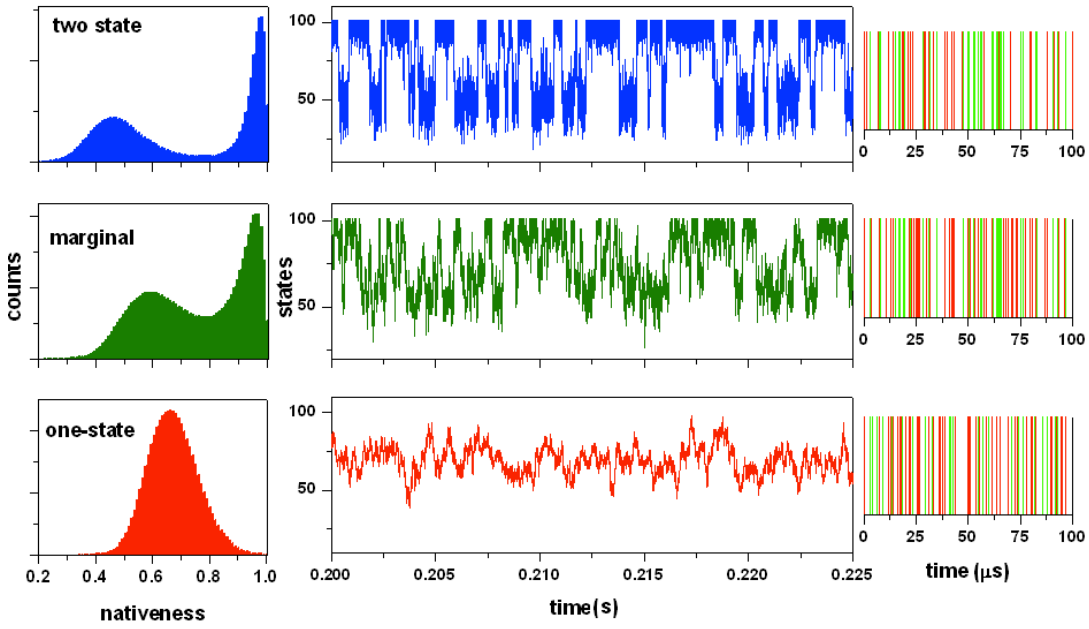


Figure 4.2 Sample Stochastic State Trajectories of 25 ms and distributions sampled in a 20s total stochastic simulations are shown for each of the scenarios. An example photon trajectory with donors as green lines and acceptors as red lines are also shown for each of the scenarios. The timescales of photon emissions are in μs whereas the sample trajectories represent dynamics in a protein with a relaxation rate of ~200 μs.

For fast folding proteins with rapid conformational dynamics the limitations of methods that involve binning the collected photon data for analysis is demonstrated in Figure 4.3 that shows the broadening in the FEH when binning times are much longer compared to the timescale of dynamics of the studied process. For the three different folding scenarios, 50,000 bursts from free diffusion experiments were simulated. As the dynamics has been calibrated with a relaxation time of ~200 μs, the binning effects on the FEH are revealed by using different binning times (50, 200 and 1000 μs) with appropriate thresholds. When the binning times are shorter than the relaxation time, distinguishable bimodal distributions are observed in the FEH for both the two-state and marginal barrier scenarios and a singular unimodal distribution is observed for downhill scenario. In the case of being able to have sufficient photon counts within the relaxation times, FEH could serve for distinguishing subpopulations of the different states in the system by showing multimodal distributions or unequivocally point to a single state, after appropriately accounting for the effects of shot noise.[100]. But as the bin times becomes comparable to the relaxation times, the resolution of subpopulations is progressively difficult with gradual merging of the peaks even for two-state systems. When the bin times are much longer compared to the relaxation times, dynamic averaging happens and the resulting FEH are broad and unimodal. In such cases, there is no way to identify if the underlying scenario is truly one-state or has more states than one that are separated by a barrier. This becomes an important problem especially for fast-folding proteins that have dynamics in the μs regimes and collecting sufficient amount of photons within their relaxation times is not possible. In such cases, binning free methods that directly utilize all the information available in the photon trajectories are then necessary for pushing the time resolution of the

smFRET experiments for studying such fast-folding proteins and obtaining their dynamics.

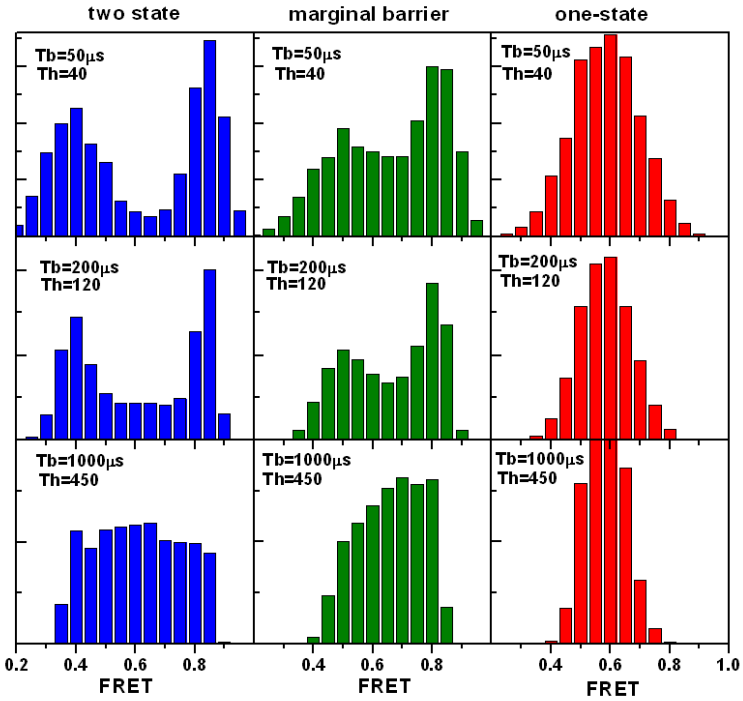


Figure 4.3 Sample (100 μs) Photon Arrival Trajectories for the 3 different scenarios and FRET Efficiency Histograms (FEH) with different Binning Times of 50, 200 and 1000 μs with corresponding thresholds of 40, 120 and 450 from 50,000 simulated bursts of 250 μs average residence times

4.3.1 Parameter recovery by the procedure and testing its robustness

We apply the maximum likelihood analysis (MLA) procedure on simulated photon arrival trajectories of three different folding scenarios (1-D FES model parameters given in Table 4.1) to identify the input model parameters from them. The advantage of using synthetic data is that the model parameters are completely known and the conformational dynamics of the protein are also available in the stochastic state trajectories. Thus the effectiveness of the procedure to recover the FES model parameters using the photon trajectories alone could be quantitatively evaluated.

Table 4.1. Results of the simple FES model combined with MLA procedure. Parameters recovered using the MLA procedure, from 50 iterations with total 100,000 photons at 800 ms^{-1}.

Scenarios	$\Delta H_{loc,res}$ (kJ/mol)		$\Delta H_{nonloc,res}$ (kJ/mol)		log10(D)	
	Real	Recovered	Real	Recovered	Real	Recovered
Two-state	2.70	2.69±0.02	3.81	3.81±0.01	7.00	6.99±0.03
Marginal Barrier	3.72	3.72±0.03	3.36	3.35±0.01	6.31	6.32±0.03
Downhill	5.42	5.49±0.12	2.31	2.26±0.08	5.51	5.49±0.02

Results from 50 different iterations of applying the procedure on simulated photon trajectories of 100,000 photons at a total countrate of 800 per millisecond for each of the folding scenarios for a protein with a size of 50 residues and at a temperature of 298K is shown in Table 4.1. For each of the iterations, an independent stochastic simulation was done for generating simultaneously generating both the state and photon arrival trajectories as described in the Methods, followed by application of the combined procedure on the photon trajectories. For all the three different simulated scenarios, the procedure recovers input model parameters with high accuracy. For the scenarios with barriers, the parameters are more specific and determining of the exact height of the barrier. In the case of downhill scenario with zero barriers, a range of parameter values produces similar 1d- free energy profiles. Thus the parameters $\Delta H_{\mathrm{loc,res}}$, $\Delta H_{\mathrm{nonloc,res}}$ identified by the maximum likelihood procedure need not be exact for downhill scenario but still reproduce the zero barrier 1D-FES. Accurate determination of the dynamic parameter D is important though as the parameter for determining the timescale of the folding process. From Table 4.1, it is evident that the dynamic parameter D is reliably recovered for all the scenarios and with lesser errors in the case of downhill scenario.

Photon trajectories typically measured in the experiments are with varying photon countrates and are almost always with background photons either in acceptor or the donor channels or mostly in both the channels. Thus it is crucial to evaluate the procedure using input photon trajectory data of varying quality and content. We perform the following tests Case A) Changing the total amount of data used in the procedure i.e. different number of photons while maintaining the same total photon count rate, Case B) Varying photon count rates while maintaining the same total number of photons in the trajectories and Case C) Adding background noise to the photon trajectories.

4.3.2 Case A) Dependence on Data Availability

When the total countrate is the same for all the conformational states and does not vary from state to state, the maximum likelihood procedure is rigorous and is able to robustly distinguish between the folding scenarios and identify the right dynamics (model parameters) even from very short trajectories. To test this, we varied the total amount of data from 5000 to 100000 photons while maintaining the photon countrate constant at 800 ms^{-1}. The procedure correctly identified the folding scenarios even with short photon trajectories as inputs. From the identified (fit) FES model parameters, probabilities were computed and compared to those computed with the true model parameters. For making comparisons with the conformational dynamics of the molecule within the short time lengths of fewer photons, we compute normalized frequency counts of the microstates from the corresponding stochastic conformational trajectories (referred as simulation input). Figure 4.4 shows the comparisons for three different lengths of trajectories for each of the scenarios. Clearly, the system samples only a subspace of conformations when the photon strips are very short as shown by the black curves (input simulations). The recovered probabilities

are closer to the parent or true probabilities in all these cases even with short input trajectories of just 10,000 photons.

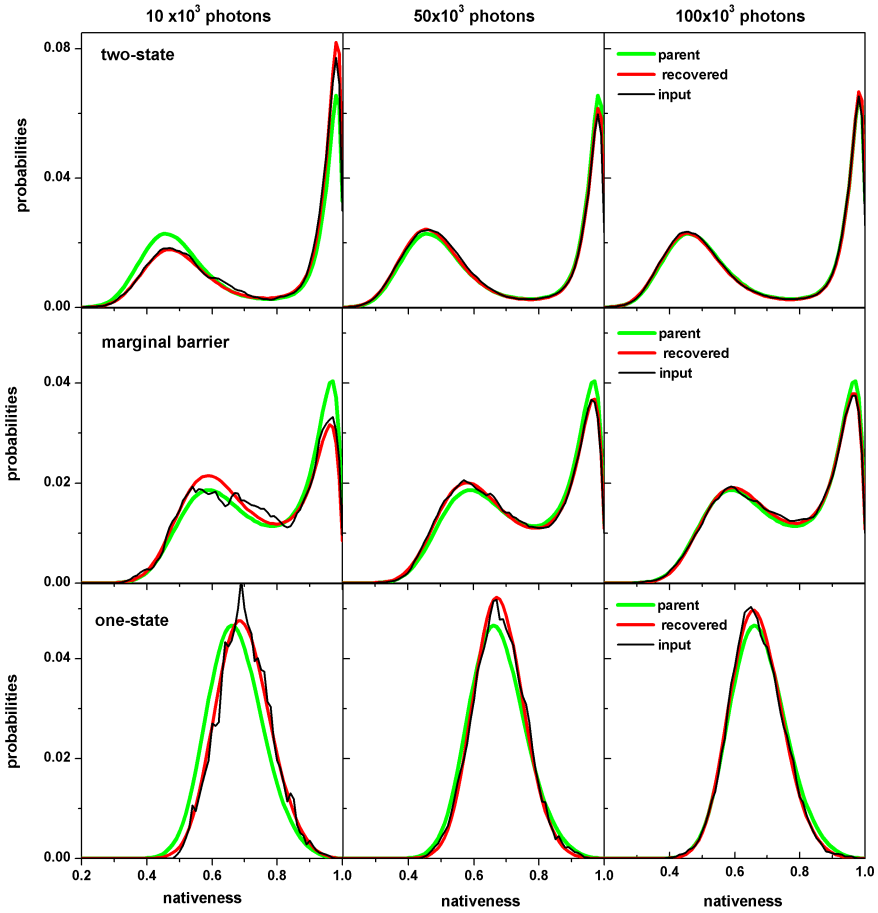


Figure 4.4 Dependence of the procedure on the amount of available data. Here we show the recovered probabilities, parent profiles and the normalized counts of the input simulations for different amount of data (number of photons) for each of the three scenarios. The procedure performs well already at 5000 photons and the accuracy increases with more data.

% deviation between the input probabilities from the simulations and the recovered probabilities using the MLA procedure are shown in Figure 4.5. For comparing the dynamics parameter an RMSD score is computed between the recovered and the input parameter D (in log space as log10(D)).

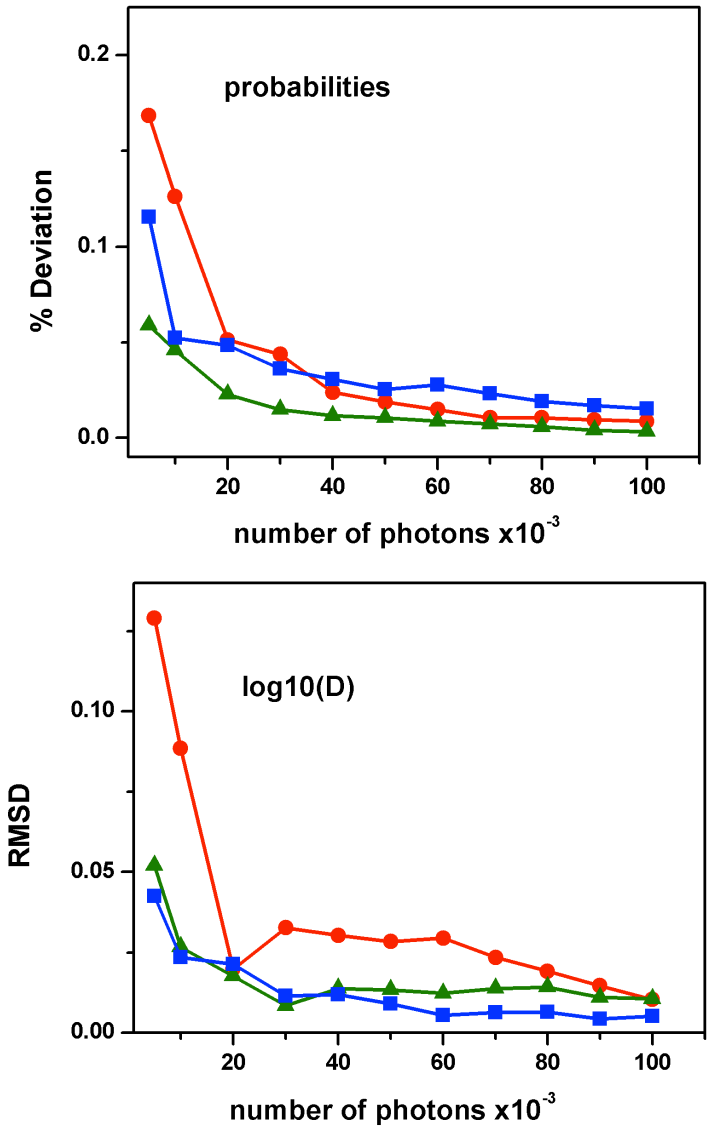


Figure 4.5 Effect of different amount of data (varying number of photons) A) Here we show the %Deviation between the recovered probabilities and the normalized counts in the input simulations vs. the total number of photons in the photon trajectories (B) RMSD of dynamic parameter 'D' which gives the dynamics of the process is shown vs. the total number of photons in the photon. (Legend: blue –two-state, green – marginal and red- downhill scenarios)

4.3.3 Case B) Effect of low total photon countrates on identification of fast conformational dynamics

Collecting more photon data relative to the relaxation times is of importance to resolve fast processes using smFRET method. Here, we investigate the performance of our combined procedure to identify the dynamics of fast folding and to particularly examine the effects of low total countrates on the efficacy of the methodology. The total countrates achievable, especially with free diffusion experiments are limited (maximum theoretical limits of 1 μs^{-1} owing to nanosecond fluorescence lifetimes of the dyes and detection efficiencies as mentioned earlier) and are also dependent on the experimental systems. As

mentioned earlier, for fast folding proteins with conformational dynamics in the sub-microsecond timescales, this limitation is one of the main drivers for development of such new and powerful methods for photon-by-photon analysis of the data rather than the traditional methods that involve binning of data. Using this test, we establish that the procedure robustly identifies the FES model parameters; in particular dynamics even when only small photon count rates are available. This test is directly related to the experimental parameters for achieving an optimal count rate without incurring photodamage, bleaching and still being able to identify dynamics of the protein from it. That is, the experiment could be so designed so as to not use very high laser intensities that is the main cause for the photo damage of the fluorescent dyes and even with a bit lower count rates achievable without such measures, the dynamics of interest could be deciphered from the data.

We varied the photon count rates (from 10 to 800 ms^{-1}) keeping the total number of photons for the analysis constant at 100,000 photons. With lower count rates, for achieving the same total number of photons the measurement time has to be much longer. In our simulations, the simulation length (number of steps) has to be increased thus extending the conformational sampling in the state transition trajectories and the total measurement time. As the timescales of the conformational dynamics, τ = 200 μs for the different scenarios is known, the effect of varying photon count rates on recovery of the dynamic parameter D in particular could be evaluated from this test. Figure 4.6 shows the %deviation in the probabilities and RMSD in the parameter D versus the ratio between the average inter photon arrival times and the relaxation time (200 μs). As the dynamic information is extracted only from the photon color sequence & the inter photon arrival times, the number of photons within the relation time constant of the process of interest (fast folding) is the most relevant determinant for the efficacy of the procedure and so we have used it as the abscissa. The method performs well even with very few photons arriving within the relaxation time of the process (at countrate as low as 1% of the relaxation rate). Thus the time resolution of smFRET experiments could be drastically improved by adopting this procedure for photon-by-photon analysis.

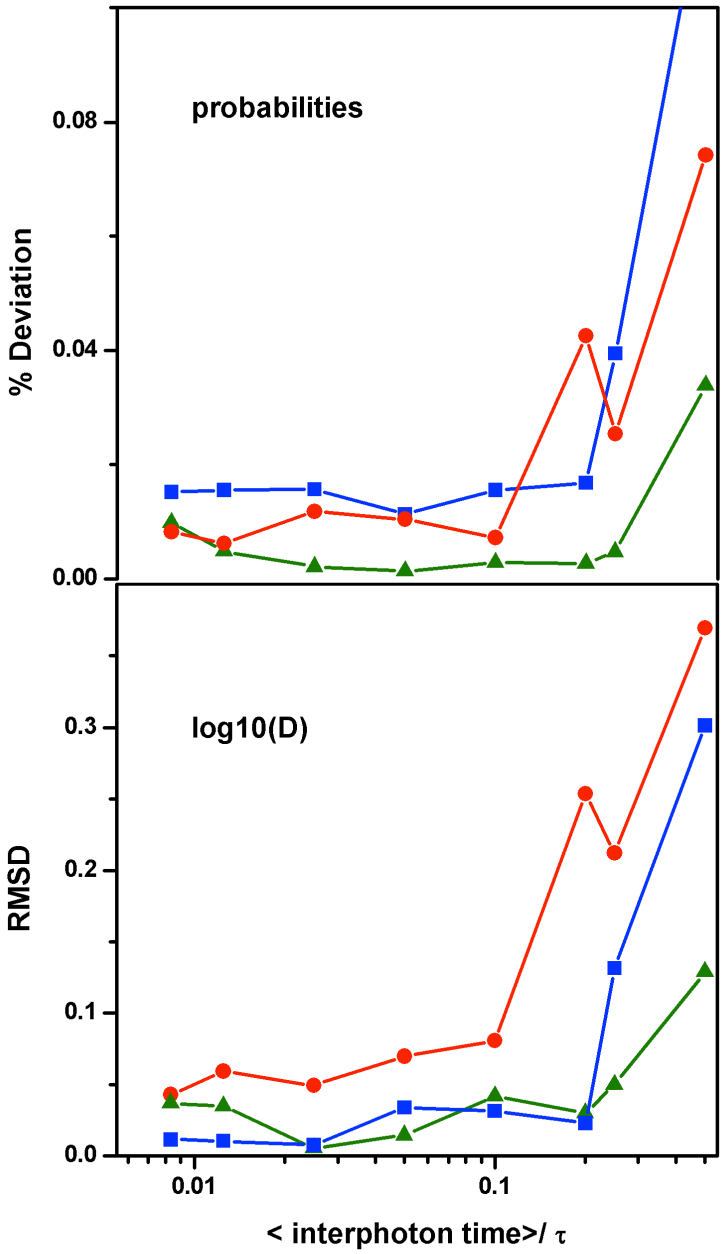


Figure 4.6 Effect of varying photon count rates, taking same quantity of photons (100,000). Total duration of photon trajectories and the total sampling in the state trajectories required to generate them depends on the photon count rate. A) Here we show the %deviation between the recovered probabilities and the normalized counts in the input simulations vs. the ratio between the average inter photon arrival times of photons and the relaxation time τ = 200 µs (B) RMSD of dynamic parameter 'D' which gives the dynamics of the process is shown relative to the ratio between the average inter photon arrival times of photons and the relaxation time τ = 200 µs. (Legend: blue –two-state, green – marginal and red- downhill scenarios)

4.3.4 Case C) Effect of Background Noise

The above tests were performed with noiseless trajectories but experimental data invariably contains background photons detected in both the donor and acceptor channels due to various effects resulting from photophysics or

instrumentation. The typical effect of such background noise is modifying the shape of the FEH. The maximum likelihood procedure itself is robust with the presence of background noise and in particular the dynamics are identified correctly**.** An example for the effect of background photons on FEH and the ability of the procedure to identify the right scenario even from noisy data is shown in Figure 4.7. In this case, for the marginal barrier scenario, equal amount of noise has been added to both the donor and acceptor channels and the FEH with 10% noise is shown as an overlay on the FEH from noiseless trajectories of 100 ms. Effect of noise is seen on the dynamic range of the FEH. When the background noise is present only in one of the channels, FEH simply gets shifted in their peaks accordingly whereas the presence of similar amount of noise in both the acceptor and donor channels leads to compression in the FEH indicating a reduced dynamic range of the peak FRET values. Result of the MLA procedure using noisy data as input is shown as an inset in the Figure 4.7 in which the recovered probabilities, distribution of states in the input simulations and the parent distributions are shown. Similarity in these curves, demonstrates the performance of the method in identifying the right scenario based on noisy data. With increasing noise levels, there is a drop in performance and at 10% noise levels the parameters recovered are not very accurate though still pertaining to the same folding scenario.

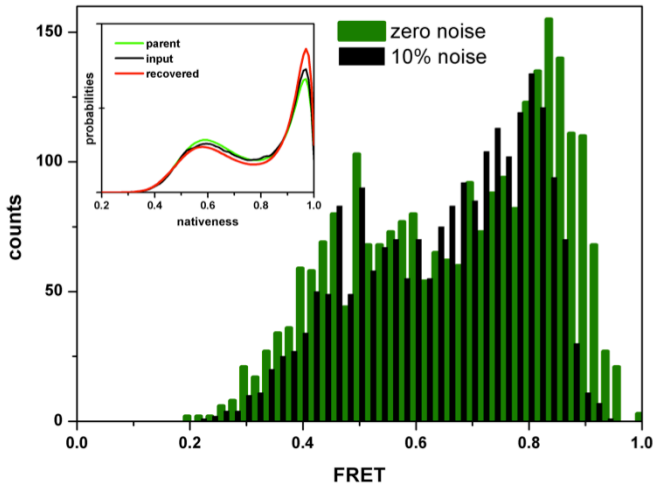


Figure 4.7 Effect of background photons on the FEH. FEH from marginal barrier scenario using a total 100ms of noiseless photon arrival trajectories and trajectories with 10% noise in both the donor and acceptor channels, done with a binning time of 200 μs is shown here. Inset shows the probability distributions of the parent, input simulations and the recovered probabilities for the 10% noise case.

For our combined procedure, since there are added restrictions in terms of mapping functional between the inter-dye distances and nativeness (order parameter), we performed tests for studying the effect of noise on its performance. In the case of background noise, the fitted FRET efficiencies are transformed with a correction to obtain the correct FRET efficiencies according to:

$$\varepsilon_i^{fit} = ((< n_t > -b_A - b_D)\varepsilon_i^{calc} + b_A)/ < n_t > \qquad (4.5)$$

where $< n_t >$ is the total countrate, b_D and b_A are the background rates in the donor and acceptor channels, ε_i^{calc} is the FRET efficiencies obtained directly from the distance mapping to the order parameter. With this correction, the combined procedure performs remarkable well for a range of noise levels. Figure 4.8 shows the results of the procedure with the % of background (extra) photons in the data as the abscissa. The downhill scenario is more robust (less sensitive) to the presence of extra photons compared to the activated & marginal barrier scenarios.

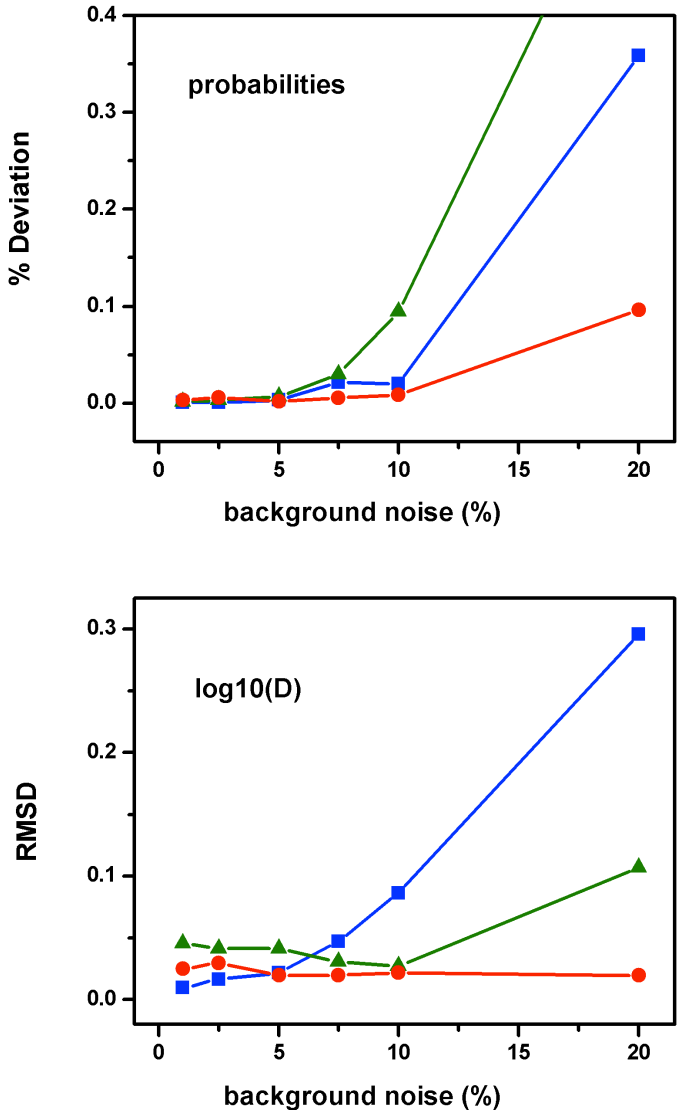


Figure 4.8 Effect of background photons on the procedure. A) Here we show the %deviation between the recovered probabilities and the normalized counts in the input simulations vs. % of background photons relative to the total number of photons in the data (B) RMSD of dynamic parameter 'D' vs % of background photons relative to the total number of photons in the data. (Legend: blue –two-state, green – marginal and red- downhill scenarios)

4.4 Concluding Remarks

Our combined procedure extends the powerful maximum likelihood method developed by Gopich and Szabo [120] to easily and directly extract relevant thermodynamic and kinetic parameters of fast folding proteins from timestamped smFRET photon trajectories. The maximum likelihood method has so far been only applied with simple two or three-state kinetic models. By implementing the likelihood method on a continuous Free Energy Surface model we have expanded its scope to directly obtain the conformational dynamics of the protein from smFRET experiments. Though we have demonstrated it here on a particular version of the simple free energy surface model, the procedure could be extended to any other model that formulates the folding process as a simple diffusion over an appropriate reaction coordinate and has the kinetics defined by a rate matrix. We have tested this by implementing the procedure on other versions of the simple free energy surface model [51] as well as on surfaces that are sums of simple harmonic potentials. Thus, the procedure is applicable for fast folding proteins with a spectrum of behavior from simple two-state kinetics or with intermediates to completely one-state. Our results also offer guidelines for experiment design by indicating that the dynamics could be correctly identified even with small quantity of photons when the count rates are sufficiently higher as compared to the dynamics of the folding process. The presence of background photons is unavoidable but to have high accuracies care must be taken by use of appropriate filters to reduce the noise in the channels. Some other typical concerns in the experiments such as dye "blinking" can be addressed by using an appropriate kinetic model in which the dark states are included in the likelihood function itself. With successful application of the simple free energy surface model of the protein folding to analyze and understand complex data such as smFRET photon trajectories, we have yet again demonstrated the applicability and power of such simple free energy surface models.

Part II: Statistics in Protein Folding

5 Higher Order Φ/Ψ Maps and Derivation of Entropic Costs of Protein Folding

5.1 Introduction

In 1960s, Ramachandran and Shasishekaran[124] discovered that simple steric clashes restrict the rotatable bond angles in a polypeptide chain to limited areas in space, which has come to be represented as the Φ-Ψ map or Ramachandran Plot for proteins. In a peptide bond, as the ω dihedral angle is typically planar and fixed (to 180º) only the φ, ψ angles are rotatable. When a polypeptide chain folds, consecutive monomer units of the chain linked by the planar peptide bond are limited to occupy certain regions in the dihedrals space due to steric clashes. Secondary structures in proteins have characteristic patterns of preferred dihedral angles that could be used to define them. That is, segments in the polypeptide chain having particular values of dihedral angles are called as alpha helices or beta sheets etc. accordingly. Interestingly, the existence of such regular arrangements in protein backbones were first predicted theoretically based on hydrogen bonding patterns[125] and were later confirmed with experimental determination of protein structures using X-ray crystallography and subsequently Nuclear Magnetic Resonance (NMR) and other methods. Steric restriction, as identified by Ramachandran et al. using simple hard sphere approximations for the atoms is another determining factor for these regular secondary structure elements. Ramachandran plot has been a very useful tool since then to evaluate protein structures and has become a standard in protein structure elucidation and validation using any of the experimental structure determination techniques. All structure validation programs now routinely check for the fraction of the amino acids in unfavorable Ramachandran regions[126] and rank protein structures as good or bad based on this fraction among many other parameters.

Torsional angles in general and the φ/ψ backbone dihedral angles in particular also form an important facet of the simulations and analysis of protein folding and protein structures. Firstly, as angular coordinates they offer a reduced degree of representation of a polypeptide chain compared to the 3-D Cartesian coordinates of the atoms in the polypeptide. This simplified representation is used in many molecular mechanics (MM) and molecular dynamics (MD) simulations of proteins in order to speed up the calculations by an order of magnitude by performing moves in the torsional angle space rather than in the regular Cartesian space. For example, most structure calculation programs used in NMR perform the dynamics of the protein in torsion angle space[127]. Secondly, for analyzing the large amounts of data generated from MD simulation, the dihedral angles are very useful again by offering a compact representation of the protein structure. Conformational sampling in the simulations is presented as the sampling of particular regions of the Ramachandran plot and the side chain

dihedral distributions are also often presented to visualize the conformational dynamics of particular amino acid chains in the MD simulations.

The importance of the backbone dihedral angles stems from the fact that they are mainly a "local" characteristic of a protein backbone. According to the Energy landscape theory of protein folding the multi-dimensional complex energy landscapes of proteins could be projected down into one or two dimensions by using appropriate order parameters as reaction coordinates as we have already seen in Chapter 1. Along with the fraction native contacts, backbone dihedral angles are typically considered to be suitable coordinates for representing the free energy surfaces of proteins in reduced dimensions. In simple statistical physics based models of protein folding, such as the Ising-like simple FES model introduced earlier (in Chapter 3) the total free energy of the system is partitioned into the contributions of the individual elements and is taken as the sum of the interactions between these elements. For representing the state of the individual elements, peptide bonds or amino acids in a protein, the backbone dihedral angles form superb reaction coordinates, as they are completely 'local'. A particular peptide bond could be considered folded or unfolded based on whether it is in the native dihedral angle or not. In such simple statistical models, degree of protein folding is given by the number or fraction of units (peptide bonds or amino acids) taking the values observed in their native conformations. If only all the units are in native-like dihedrals, then the protein is fully folded. This approach was first used in the models by Zwanzig[43a] and later developed in the simple models of folding by Munoz-Eaton[44]. Being such a fundamental local reaction coordinate, the dihedral angles offer a key connection to the 'orderedness' of a polypeptide chain, which is the main topic of this research study. Our aim is in exploiting the crucial relationship between the progressive ordering of polypeptide backbone angles into their native dihedrals as used in the simple free energy surface models of protein folding and the corresponding entropic costs of protein folding. We use a simple approach to partition the Ramachandran plot into different higher order regions or clusters and use such clusters to calculate the entropic cost of fixing particular amino acids in them.

5.2 Entropy of Protein Folding

The protein folding process by definition is a process of ordering. As the protein folds from its denatured state to the native state there is a huge loss in conformational entropy since the amino acids are adopting specific and restricted backbone angles compared to their conformational freedom in the unfolded states. This entropic loss constitutes the major unfavorable contribution to the free energy of folding and needs to be overcome by favorable interaction energies for the folding process to occur. In fact, this entropic loss is the main reason for the relatively low ΔG values for typical protein folding, which is a process with a compensation between the ΔH and ΔS, referred to as Enthalpy-Entropy compensation. Free energy barriers to folding are thought to arise out of the asynchrony in this compensation, which is a direct implication from the Energy landscape theory (Chapter 1). As the folding progresses, initially the entropy losses are much higher than the gains in interaction energies

and with the progress of the process, enough energy is gained so as to nullify and overcome the losses in entropy. The point with the maximum differential between the entropic loss and energetic gains is the top of the folding barrier. In spite of the importance of entropic factors in the energetics and kinetics of protein folding proper evaluation of folding entropies has been difficult and direct experimental measurement of the entropy of unfolded state elusive.

Efforts to estimate the entropy of unfolded state have been through the following approaches:

1) Purely theoretical or computational approaches by estimating the number of possible conformations accessible for the amino acids
2) Semi-empirical approaches that have experimentally measured total entropic changes for protein folding and subtracting out contributions other than conformational contributions.
3) Empirical estimates based on experiments such as model-dependent direct estimates from NMR relaxation experiments, neutron scattering experiments, and estimation from pulling experiments using Atomic Force Microscopy

Computational approaches for estimating conformational entropy of folding by enumerating the number of states accessible in unfolded states have been attempted since many decades [128]. Pauling and Corey in 1951 estimated 72 conformations per residue converting to a backbone entropy cost of 36 J $mol^{-1}K^{-1}$ [125]. This was reviewed by Schellman in 1955 who estimated the range to be between 12 J $mol^{-1}K^{-1}$ and 30 J $mol^{-1}K^{-1}$ per residue[129]. Later in 1965 using computer simulations, Nemethy and Scheraga estimated the number of conformations of Glycine as 21 and non-Glycine amino acids to be 7 amounting to entropic costs of 25 J $mol^{-1}K^{-1}$ and 16 J $mol^{-1}K^{-1}$ respectively given all the conformations are populated equally[130]. In the 1990s, Wang and Purisma employed detailed simulations to estimate a cost of 21 J $mol^{-1}K^{-1}$ for folding a residue at the center of an alpha helix[131]. In 1995, Freire and co-workers combined computer simulations and mutational experiments to derive a scale of entropy values for different amino acids with a mean of 14.52 J $mol^{-1}K^{-1}$ and loss of 10.1 J $mol^{-1}K^{-1}$ going from Glycine to Alanine with the introduction of the methyl side chain[132]. Leach et al in 1960s evaluated this loss of adding a methyl side chain to Glycine leads to a 3.25 factor reduction in the number of accessible conformations resulting in a conformational entropy difference of 9.64 J $mol^{-1}K^{-1}$. By counting the number of rotamers observed in protein structures, estimates of 18 J $mol^{-1}K^{-1}$ has been proposed by different researchers[133,134] . Using semi-empirical approaches of subtracting out other contributions such as entropy of hydration from experimentally measured total entropies, Privalov et al, made an estimate of ~15 J $mol^{-1}K^{-1}$ per residue at 393K[135]. Zhang et al. used a similar but computational decomposition procedure to arrive at an estimate of 22 J $mol^{-1}K^{-1}$ at room temperature[136]. Assuming simple models of bond vector motions, some research groups have estimated the backbone entropic costs by correlating with NMR order parameters obtained from relaxation data on model proteins. It ranges from 15-20 J $mol^{-1}K^{-1}$ for residues totally disordered in denatured states and a lower 6-12 J $mol^{-1}K^{-1}$ for residues retaining some residual order even in

denatured states, as estimated by Kay and coworkers[137]. Alexandrescu and coworkers, working on S-peptide folding estimate costs of 13 J $mol^{-1}K^{-1}$, 18 J $mol^{-1}K^{-1}$ or 23 J $mol^{-1}K^{-1}$ of entropic loss for residues fully disordered in the denatured states, depending on the models employed to correlate bond vector motions with entropy[138]. Using neutron spectroscopy measurements and models for correlating scattering dynamics with entropy, Jörg Fitter estimated 10.85 J $mol^{-1}K^{-1}$ per residue at 303K[139]. Measuring entropy lost when pulling mechanistically stable beta proteins using atomic force microscopy (AFM), Plaxco et al., in 2002 made an estimate of 19±2 J $mol^{-1}K^{-1}$ per residue [140].

Entropic changes during folding have two components – the conformational entropy (ΔS^{conf}) which is primarily due to the loss of configurational freedom and the solvation entropy ΔS^{solv} (or hydration entropy). The solvation entropy could be further divided into polar and nonpolar components arising from the burial of polar and non-polar groups. As the polar component, ΔS^{polar} is negligible and has been shown to be zero around 335K, the solvation entropy has been deemed significant only for non-polar groups ($\Delta S^{non\text{-}polar}$) whose solvation gives rise to the hydrophobic effect[141,142]. Using transfer of model compounds to water from non-polar solvents, these non-polar entropies have been measured to be negative at room temperature and to be proportional to their water accessible surface areas (ASA). They are temperature dependent with their magnitudes decreasing with increases in temperatures and found to be zero and vanishing at about 112° C. Indeed as early as 1974 Privalov[143] noticed that both the entropy and enthalpy changes in protein folding when normalized by the number of residues, show convergence at higher characteristic temperatures of around 100˚C. Murphy and Gill, based on comparative thermodynamic studies between proteins and model compound transfer studies found this convergence temperature corresponds to the temperatures at which nonpolar contributions become zero[144]. For the entropies, this convergence temperature was located to be around 112˚C the same temperature where the model compound $\Delta S^{non\text{-}polar}$ becomes zero, whereas it remained to be at 100˚C for enthalpies. Freire and coworkers in 1992[145] confirmed this with further studies and pinned the origin of the convergence to hydrophobic effect as proposed by Baldwin in 1986[142b]. Besides the thermodynamic data, the convergence temperature has also been confirmed based on protein folding kinetics data by Akmal & Munoz, in 2004[146]. Though there has been fair amount of debate over the convergence temperatures and the proposed basis for this convergence, it has been accepted that the contributions from solvation is negligible at these temperatures. Thus, ΔS values for protein folding extrapolated to 385K, which is the convergence temperature, has only contributions from the conformational entropy ΔS^{conf}.

5.3 Experimental characterization of thermodynamics of protein folding: a background

Stability of protein molecules is quantified by the Gibbs free energy ΔG_u between the folded and the unfolded states. The totally denatured states reached upon perturbations such as temperature, pH or chemical denaturants such as guanidinium chloride (GdmCl) or urea are considered to be the unfolded states. Though there are speculations of some residual structures, differing extents of

hydration etc. in the denatured states and differences between the denatured states obtained with different perturbations, as far the thermodynamics of the process, the states reached on denaturation with the perturbation such as temperature are operationally defined as the unfolded states.

Equilibrium constant between the native and denatured states is defined as:

$$K_{eq} = \frac{[D]}{[N]} \qquad (5.1)$$

and this holds regardless of the presence or absence of intermediates between these end states.

From the equilibrium constant, the free energy is given as,

$$\Delta G_u = -RT \ln(K_{eq}) \qquad (5.2)$$

where R is the universal gas constant and T is the absolute temperature. As the free energy change is temperature dependent, from the Gibbs relationship

$$\Delta G_u(T) = \Delta H_u(T) - T\,\Delta S_u(T) \qquad (5.3)$$

where ΔH_u and ΔS_u are the enthaplic and entropic differences between the states at the same temperature, T. Both the ΔH_u and ΔS_u are also dependent on the temperature and this dependence is determined by ΔC_p, the heat capacity change between the native and the unfolded states. Unfolded states have higher heat capacities compared to the folded states of proteins because of their increased degrees of freedom and increased solvent restructuring around them i.e. the energy storage capacity of the unfolded state is higher. So, the amount of heat required to raise the temperature of a solution of protein in their unfolded states is always greater than required for a corresponding raise of the temperature of the solution when the protein is folded. While there is a slight dependence of ΔC_p on the temperature, assumption of constant ΔC_p typically holds and does not result in significant error in the calculation of any other thermodynamic parameters. Taking a convenient reference temperature Tr, the dependence of ΔH_u and ΔS_u is given by

$$\Delta H_u(T) = \Delta H_u(T_r) + \Delta Cp\,(T - T_r) \qquad (5.4)$$

$$\Delta S_u(T) = \Delta S_u(T_r) + \Delta Cp \ln (T/T_r) \qquad (5.5)$$

With this temperature dependence of ΔH_u and ΔS_u, ΔG_u as a function of T could now be expanded as:

$$\Delta G_u(T) = [\Delta H_u(T_r) + \Delta Cp\,(T-T_r)\,] - T[\Delta S_u(T_r) + \Delta Cp \ln (T/T_r)] \qquad (5.6)$$

$$\Delta G_u(T) = \Delta H_u(T_r) - T\Delta S_u(T_r) + \Delta Cp\,[\,(T-T_r) - T \ln (T/T_r)\,] \qquad (5.7)$$

If Tm, the midpoint temperature of thermal denaturation is taken to be reference temperature T_r, then

$$\Delta G_u(T = Tm) = 0 \qquad (5.8)$$

$$\Delta Sm = \Delta Hm / Tm \qquad (5.9)$$

where ΔH_m is $\Delta H_u(T = Tm)$ and $\Delta S_m = \Delta S_u(T = Tm)$
Then,

$$\Delta G_u (T) = \Delta Hm (1 - T/Tm) + \Delta Cp [(T - Tm) - T \ln (T/Tm)] \qquad (5.10)$$

which is referred to as modified Gibbs-Helmoltz equation, incorporating the temperature dependence of ΔG_u

Dividing equation 5.10 by –RT and applying equation 5.2 leads to,

$$\ln K = 1/RT [\Delta Hm (T/Tm -1) - \Delta Cp [(T-Tm) - T \ln (T/Tm)]] \qquad (5.11)$$

Experimental values of lnK at different temperatures could then be fitted to yield values for the parameters Tm, ΔH_m and ΔCp, from which ΔS_m could also be directly obtained. In this case the crucial assumption is that there are no stable intermediates in the process of F <-> U, and the experimentally measured values of K are a true measure of the equilibrium constant of the transition.

5.3.1 Differential Scanning Calorimetry (DSC)

Differential Scanning Calorimetry (DSC) is a powerful technique for characterizing the thermodynamics of conformational transitions in proteins, particularly the global folding to unfolding transitions. Using DSC, the partial molar heat capacity of a solution is determined as a function of temperature. A typical DSC instrument has two cells, one for the protein solution and one for the buffer solution and has a feedback mechanism so as to maintain both these cells at same temperatures i.e. at zero temperature difference. The cells are heated at a constant rate called as the scanning rate that is typically ~0.5 to 1 K per minute while maintaining the zero differential in temperatures. Since the protein and buffer solutions have different heat capacities, there is a difference in power requirement for maintaining the two cells at same temperatures. The ratio of this power difference of the heat flow (J/s) to the scanning rate (K/s) then directly gives the amount of heat supplied for the raise in temperature i.e. the heat capacity of the system. Since it really is a differential between the protein solution and just the buffer, $\Delta Cp^{app} = \Delta Cp^{protein} - \Delta Cp^{buffer}$ some simple mathematical manipulations are needed for deriving the partial molar heat capacity of the protein. (Eq. 5.12)

$$Cp^{protein} = \Delta Cp^{app} / C. V_o. 10^{-6} + (V_{prot} / V_{solv}) \Delta Cp^{buffer} \qquad (5.12)$$

where C is the concentration of the protein in millimolar units, V_o is the volume of the calorimetric cells in milliliters, V_{prot} and V_{solv} are molar volumes of protein and solvent buffer respectively, which are standard values obtained from the literature. $Cp^{protein}$ is the partial molar heat capacity of the protein referred to as <Cp>. The excess heat capacity for the protein solution is due to protein molecules in the solution undergoing a temperature dependent conformational change and in the unfolded state they have higher heat capacities, as mentioned above. Measuring <Cp> is non-trivial as the technique is very sensitive and many factors such as tiny errors in concentrations of protein in solution could lead to errors in the measured thermodynamic parameters. The <Cp> function (endotherm) could be analyzed to provide all the necessary thermodynamic data about the conformation change being studied. The area under the <Cp> function yields the ΔH_m of transition, as heat capacity is the derivative of enthalpy with respect to temperature. The shift in the baselines (pre and post transition baselines) gives the value for ΔCp, the difference in heat capacities between the folded and unfolded states. The DSC thermogram gives direct access to the partition function of the system under study and thus all the required thermodynamic information could be directly extracted from it.

Though a powerful technique, there are practical issues limiting its applications. For one, larger quantities of proteins (0.5-2 ml at concentrations of 0.5-1 mg/ml) are required for the measurements that necessitate purifying copious quantities of the protein of study, unlike other experiments such as spectroscopy. The propensities of proteins to aggregate under such higher concentrations for example of the denatured state or with self-assembly of the native proteins under high concentrations etc. also need to be addressed. Thus an assessment of dependence of the determined thermodynamics on the protein concentration is a requirement. Repeatability of the measurements needs to be ensured after every experiment by performing at least two scans on each protein to check if similar endotherms are obtained. But this repeatability doesn't necessarily imply thermodynamic reversibility, a primary assumption for applying the technique and which is demonstrated by showing that the difference in scan rates does not affect <Cp> functional.

DSC is routinely applied and has been used to characterize the thermodynamics of unfolding of many proteins. In the dataset used in this study, many of the protein thermodynamic parameters have been determined using this technique.

5.3.2 Spectroscopic techniques

Different spectroscopic techniques such as CD, Fluorescence, and FTIR could be used to study the temperature or chemical denaturant induced unfolding of proteins. By measuring particular signals at different levels of these perturbations and analyzing the resulting data with particular models for the transitions one could extract the thermodynamics.

Fluorescence spectroscopy measures the perturbation-induced changes in fluorescence of either an intrinsic probe such as tryptophan or tyrosine or an

extrinsic dye attached to the protein. Circular Dichroism (CD) measures the changes in the optical rotation of circularly polarized light by the chiral elements in the proteins under different conditions. Fourier Transform Infrared Spectroscopy (FTIR) measures the characteristic IR absorption spectra of the protein molecule and could be used to follow the changes of these properties upon perturbation. Any of these techniques typically result in a spectroscopic signal that monitors the conformational transition of protein from folded to denatured states.

By fitting the resulting signal data to appropriate models of protein folding, typically a two-state model (see Chapter 2), the thermodynamic parameters for the transition could be obtained. The baselines (appropriate signals of the end states) are crucial in the determination of thermodynamic parameters using these methods. Since the baseline values are extrapolated into the observable transition zone and the fractional contributions to the signals in this region are taken from the baseline signals, the determined thermodynamic parameters are sensitive to the baselines.

5.4 Curated dataset of experimentally determined protein thermodynamic parameters

In 1997, Robertson and Murphy[147] compiled a set of experimental thermodynamic data on proteins and performed regression analysis to establish connections between coarse features of protein structures such as size, solvent accessible surface areas etc. and its thermodynamics. Relationships between the thermodynamic parameters such as ΔH, ΔS and ΔCp from calorimetric and spectroscopic studies and the corresponding X-ray and NMR derived protein structures were identified based on the regression. Databases such as Protherm, a repository for collecting experimental protein folding thermodynamic parameters from different research groups have been developed since then. For the purposes of calibrating our method for calculating conformational entropies, we needed a carefully curated and reliable set of data. Rather than the quantity, the quality of the dataset was paramount for the method. We chose the Robertson and Murphy collection to be the basis dataset. We further refined their original dataset by removing some of the proteins that are non-globular or are dimers or having large errors in the experimental measurements. By exploring the original experiments and recalculating the thermodynamic parameters, we have produced a dataset that is highly reliable. We also replaced the structures referred to in the original dataset with higher resolution versions of the same, as available in the Protein Data Bank (PDB). This resulted in a dataset of 46 proteins compared to 53 proteins in the Robertson and Murphy dataset.

5.5 Research Objectives

Developing a robust method for evaluating conformational entropies of protein folding based on statistical analysis of the backbone dihedral angle distributions of high quality protein structures and side-chain contributions from rotameric libraries. The method is calibrated and benchmarked on a reliable dataset of

experimental protein thermodynamic data.

Applying the method for determining folding entropies from 3D structures of proteins and using the values for protein specific parameterization of the entropic terms in simple free energy surface models.

5.6 Materials and Methods

5.6.1 Φ/Ψ dihedral angle calculations from dataset of high quality

A non-redundant dataset of high resolution protein structures is selected from the PDB database using the PISCES server[148] with a cutoff of resolution better than 2.0Å, less than 30% sequence identity and a R-value <0.3, which includes 8069 domains, with structures determined using X-ray crystallography and NMR methods. In the case of NMR structures, only the first model is considered in the dataset. The structures are further refined to remove non-standard and ambiguous amino acids to finally yield about 1.5 million amino acids for statistical analysis.

The backbone dihedral angles (Φ, Ψ) are then calculated using standard definitions and the dihedrals space is represented as the Ramachandran plot. This space is discretized into 3.6^{o} bins in both (Φ, Ψ) to generate a 100x100 matrix. The indices (i,j) of the matrix are used to access the corresponding regions of the dihedrals space. For example, index (1,1) corresponds to the region of the Φ-Ψ space including all the values from -180^{0} to -176.4^{0} in Φ and Ψ. The number of hits (ni) for each of the amino acids in each of these 10000 discretized bins is calculated and stored in the 100x100 matrix. Logarithm of the number of hits in the discretized 100x100 bins is used in the clustering described below.

5.6.2 Clustering of Φ/Ψ Dataset

Clustering analysis of data is a task that seeks to identify homogenous groups of objects in the dataset according to defined properties or attributes and has been applied in statistics for a long time. The objects within the identified clusters or groups are more similar or homogenous to other within the same group compared to the ones outside of the cluster or group. In this case, the dataset is that of φ/ψ dihedral angles derived from high resolution protein structures and they are to be grouped based on the following attributes: 1) the 2-D spatial relationship between them (traditional Ramachandran plot) and 2) the number of hits within a defined region of this 2-D plot which could be taken as the 3rd dimension (the height). As mentioned above, the φ/ψ space was discretized into 100x100 regions and the clustering is performed on the number of hits and the range of angular values for each of these regions. Though there are a number of algorithms developed for clustering analysis, the k-means algorithm was chosen, as it is one of most standard and in this case very relevant algorithm.

5.6.3 k-means algorithm

k-means algorithm is a popular and straightforward algorithm for partitioning n multivariate data entities into k clusters. The partitioning is performed by optimizing the mean distances between different data points and points referred to as centroids that define the centers of each of the k clusters. The procedure is an iterative refinement. Naturally, the algorithm needs an effective definition for the distance metric that is crucial for its performance. It is the partitioning of n entities into k clusters (Ci = 1,2…k-1, k) by minimizing the within cluster square distances defined as:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \qquad (5.13)$$

where the term $\left\| x_i^j - c_j \right\|$ is the distance between a data point and the cluster centroid c_j.

The algorithm is given as follows:

1) Define an initial group of centroids. This step could be done using multiple strategies. A group of k points could be randomly assigned initially as the centers of each of the k clusters, which is the most commonly used approach. k different points from the dataset could also be simply selected as the initial centroids.
2) Assign each data point or entity to one of the k clusters to the centroid of which it is the closest. This entails calculating the distances defined according to a metric that is typically Euclidean, between all the entities and each of the k centroids with simple minimum square distance criteria assigning each of entity to a cluster.
3) Recalculate the centroids i.e. the centers of each of the k-clusters. The values of the centroids are updated by calculating the mean of all the entities that has been assigned to that cluster.
4) Iteratively repeat the above 2 steps until there are no reassignments of any of the entities are possible and the data points do not change groups.

In the case of dihedral angles, an important factor needs to be taken into consideration. These data entities form a continuous distribution and are circular by nature. i.e. the distance between -179º and 179º are not a direct Euclidean measure and these angles are closer due to the circular nature of dihedral angles. On a 360º rotation, the angles come over a full circle and thus are closer that needs to be taken into account while clustering these entities. It has been shown that the φ/ψ angle distribution indeed forms a continuous torus surface in 3-dimensions[149]. The distance metric is appropriately defined to reflect the circular nature of the dihedral angles by introducing a simple transformation in the distance formula, defined as:

$$dis\tan ce(x,y) = \sqrt{\begin{cases}(x_i - y_i)^2, & |x_i - y_i| \le 180 \\ (360 - (x_i - y_i))^2, & otherwise \end{cases}} \qquad (5.14)$$

$$x, y \in [-180, 180)$$

The effect of this transformation on the distance measure is a wrap around in the corners of the φ/ψ plot and the corner angles resulting in the correct closer distances than would otherwise with a linear metric, which is shown in the Figure 5.2. Thus when minimization of the squared distances is performed during the clustering procedure, with the periodic boundary condition implemented, clusters could extend and traverse around to include the dihedral angles from other sides and corners of the Ramachandran plot.

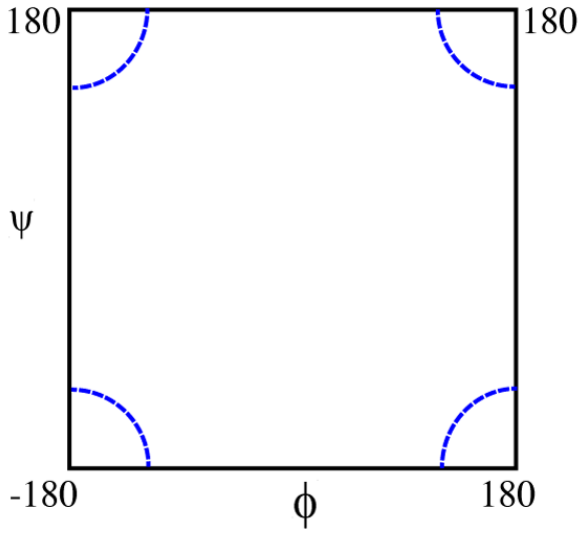


Figure 5.1 Effect of wrap around at the corners of Ramachandran Plot, illustrating the Periodic boundaries.

The primary disadvantage of k-means algorithm is the necessity of knowing the number of required clusters 'k' *a priori* as the algorithm does not offer any solution for determining this number appropriately. It just partitions the given data entities into a chosen number of clusters. The choice of k, this crucial parameter in the algorithm is user defined and thus renders the necessity of an expectation of the distribution of the data. There have been attempts to introduce dynamic determination of this parameter in other variants of k-means, but they come with their own drawbacks such as performance penalties etc. Other disadvantages include the necessity for running the procedure multiple times (replicates) and choosing the best-converged solution among them. This stems from the fact the algorithm stops at a given criteria for convergence that may be a local solution and not necessarily the global optimum along with the dependency on initial centroid assignments. Besides these disadvantages, k-means is a popular and preferred procedure owing to its simplicity and relative efficiency.

Next, the 100x100 matrix of data points is partitioned into 500 clusters using the k-means algorithm in a completely automated and non-supervised manner. As the number of clusters is data dependent and is not known *apriori,* such a large number of clusters are generated initially. In the second step, out of the 500 such partitions spatially contiguous clusters are gradually merged in a supervised manner to match the natural distribution of secondary structures typically observed in proteins. Merging of the clusters is further refined by determining the resulting entropies of folding (defined in the following section) and calibrating to their experimentally determined values. This two-step clustering procedure finally yields 10 clusters that reflect the natural distribution of φ/ψ dihedral angles and results in mean backbone entropic contributions corresponding to the experimental of ~14.6 J $mol^{-1}K^{-1}$ as calibrated to the experimental dataset. The final clusters are shown in Figure 5.3.

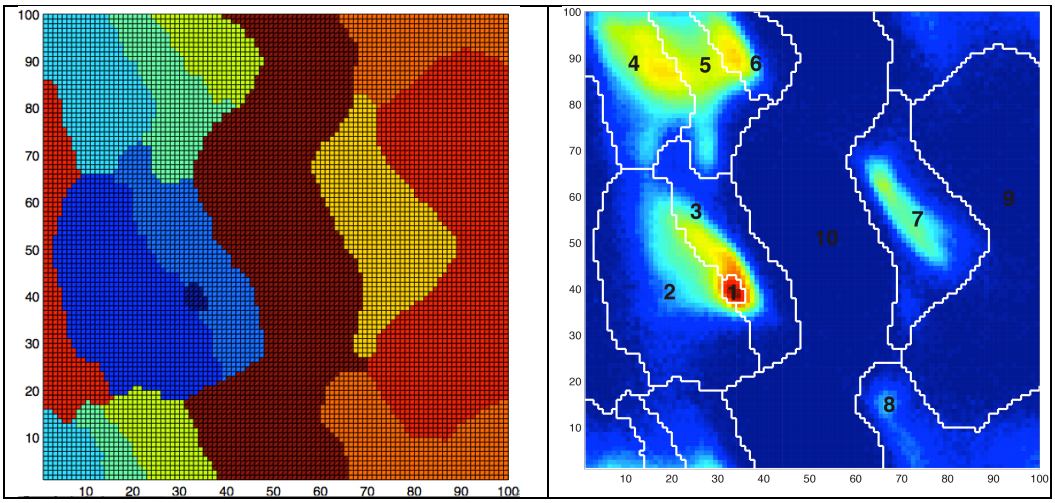


Figure 5.2 A) Final clustering with 10 clusters of the 100x100 bins; B) Cluster definition overlaid on top of the Ramachandran plot.

5.6.4 Calculation of Backbone Conformational Entropy

The distribution of backbone dihedrals in the Ramachandran space for such a large dataset can be assumed to follow a Boltzmann distribution. Based on a *microcanonical* ensemble definition of the states i.e. the amino acids having uniform probabilities to populate all the regions (any of the clusters) of the dihedral angle space, the entropic cost of fixing a given amino acid x in a cluster i is given by:

$$\Delta S_{i,x} = -R \ln\left(\frac{N_{i,x}}{N_{i,tot} - N_{i,x}}\right) \qquad (5.15)$$

where, $N_{i,x}$ is the number of hits of the amino-acid x in cluster i and $N_{i,tot}$ is the total number of hits for that amino-acid in the whole dataset. Here the amino acids are considered to populate two distinct thermodynamic states – the cluster i under consideration is taken to be native state and the rest of the regions i.e. all the other clusters correspond to the nonnative state, i.e.

$$\Delta S_i = S_{nonnative} - S_{native} \qquad (5.16)$$

From the cluster definitions given in Figure 5.4 numbers of hits of the amino acids in each of these clusters are obtained. The entropic cost of fixing each of the 20 amino acids in each of the 10 clusters is listed in Table 5.1. For any protein with known 3D structure, the total conformational entropic cost of fixing its backbone ($\Delta S^{bb,conf}$) in the native structure is then calculated as a summation of the cost of fixing each amino acid in the corresponding native cluster.

5.6.5 Calculation of side-chain Conformational Entropy

When proteins fold, the backbones adopt a very restrictive set of conformations compared to the unfolded states that contributes most to the conformational entropic cost. The side chains of the constituent amino acids in the protein chain also become more restricted due to the increased steric clashes with the ordering during folding process. From fundamental organic chemistry principles it could be easily shown that the side chains in amino acids do not possess free rotations around the C-C single bonds and adopt preferential conformations referred to as "rotamers". Side chain orientations or rotamers are specified using χ (chi) angles. According to IUPAC nomenclature, the rotamers are classified as:

Rotamer	X(degrees)
g- (gauche -)	60±60
t(trans)	180±60
g^+(gauche +)	-60±60

Depending on the number of side chain atoms, different amino acids have different number of side chain dihedral angles referred to as χ1, χ2, χ3 and χ4.

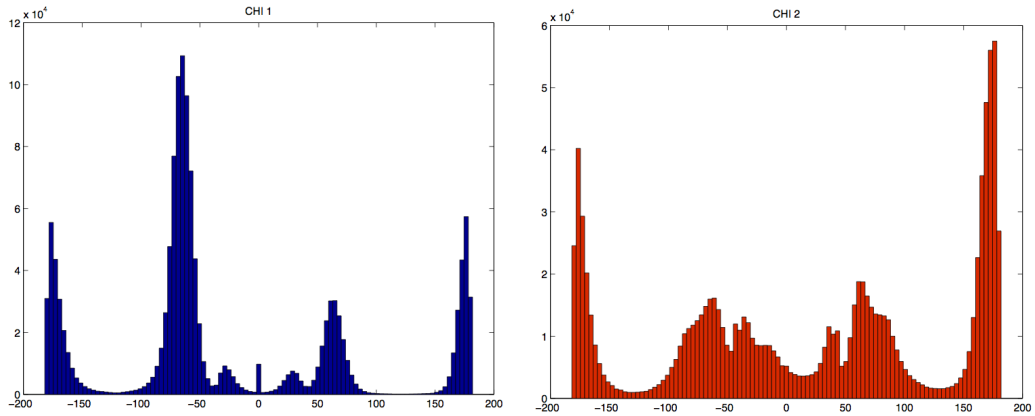


Figure 5.3 Side chain torsional angle distributions. chi1, chi2 angles in the dataset used

Studying and cataloging the frequencies of preferential side chain conformations in high quality protein structures and those sampled in molecular dynamic simulations, "rotameric libraries" have been constructed for the amino acid side chains in proteins and these libraries form a concise definition of their preferences. Such libraries have been widely used in protein modeling and

design. These could be constructed as backbone independent where the side chain torsion angles are considered independent of the mainchain dihedrals (φ,ψ) or backbone-dependent versions where the joint distributions are computed for the side chain and the main chain dihedrals. Backbone independent versions are particularly relevant for structural refinement programs in X-ray and NMR, and for developing entropic scales and representing unfolded states. Dunbrack's rotameric library[150] is one of the most commonly used and it provides the probabilities of observing different rotameric conformations for each amino acid averaged over backbone conformations observed in proteins.

Assuming a Boltzmann distribution of states, the entropy for restricting the side chains in particular conformations is calculated from the probabilities of observing these conformations

$$S = -R\sum \ln p_i \qquad (5.17)$$

where R is the Gas constant and p_i is the probability of observing a particular rotameric conformation which is taken from Dunbrack's rotameric library. We consider only the χ1 and χ2 angles for the entropy calculations as fixing these angles contribute the most to the side chain entropies. For the other angles (χ3 and χ4) the probabilities are appropriately summed into the χ1 and χ2 rotameric probabilities.

Contributions of side chains to the entropy costs of folding are not uniform as they are not restricted in similar manner upon folding. The degree of restriction of side chain upon folding depends on the number of neighbors it has in the native conformations. More the number of contacts formed by a side chain, the more conformationally restricted it is. This provides an unbiased approach to determine whether to include a particular side chain for the entropy cost calculation or not. Other measures such as residue solvent accessibility (ASA) calculations that have been taken to be indicators of the burial of a residue involve making many assumptions and are not direct, thus not preferred.

A contact is defined as a heavy atom within a distance of 4.5Å and the number of contacts is calculated for each residue from the PDB. A side chain is considered restricted or fixed if it has more than 13 heavy atom contacts. The rationale behind this particular definition is explained in following section (5.7.3). From a given PDB, both the rotameric assignments and the extent of side chain restriction could be directly evaluated without the need for any external tools or algorithms, from which the entropy costs of the side chains are derived in a straightforward manner based on the Dunbrack rotameric library.

Table 5.1 Costs of fixing each amino acid in each of the defined clusters (1-10).

	1	2	3	4	5	6	7	8	9	10
ALA	3.47	23.14	12.35	14.27	19.42	18.14	36.33	51.12	54.53	51.44
ARG	6.60	18.21	13.17	11.75	16.41	22.04	30.03	50.15	45.86	53.92
ASN	13.10	16.89	11.50	12.99	14.79	21.87	16.90	46.97	43.62	48.38
ASP	10.71	17.95	10.35	14.96	14.10	18.38	24.50	45.64	45.46	48.99
CYS	11.15	19.01	15.71	7.12	13.54	20.64	32.31	48.25	43.07	53.60
GLN	5.60	17.93	12.82	13.21	16.45	23.41	29.41	53.07	47.87	52.65
GLU	4.67	18.36	11.67	15.24	17.50	21.85	32.70	51.82	46.77	52.07
GLY	17.11	28.71	18.48	19.80	24.15	19.62	5.86	12.64	31.37	48.32
HIS	11.49	15.83	13.40	9.07	15.94	21.25	25.68	48.39	48.96	49.66
ILE	8.12	18.41	20.27	5.34	13.20	29.16	55.84	60.00	55.17	54.89
LEU	5.68	18.26	14.21	12.87	13.41	23.10	40.53	57.53	55.99	54.36
LYS	6.86	17.33	12.49	13.55	15.66	21.62	28.74	50.34	48.67	51.92
MET	6.43	19.10	13.61	11.03	15.76	22.45	34.30	51.77	40.91	51.44
PHE	10.74	17.06	14.74	7.25	13.93	23.97	34.02	55.88	57.18	54.47
PRO	19.09	39.79	6.26	60.00	18.65	0.63	60.00	60.00	60.00	50.23
SER	12.31	19.40	10.65	10.20	17.00	15.36	33.24	45.07	41.65	50.34
THR	12.21	14.62	15.57	7.94	13.82	19.59	45.84	55.46	52.73	53.26
TRP	9.17	18.82	13.02	9.84	13.57	21.93	35.42	53.45	51.60	51.78
TYR	10.90	16.61	14.68	7.19	14.47	23.64	33.61	54.33	56.87	51.91
VAL	9.83	18.66	21.42	3.72	13.35	27.23	51.03	60.00	55.58	53.86

5.6.6 Entropy costs and the Free Energy Surface Model

In the simple free energy surface models for protein folding, described in Chapter 2, entropy cost is an important parameter. The model defines an appropriate progress variable, a reaction coordinate termed nativeness that measures the different terms ΔG, ΔH and ΔS as folding progress. 'nativeness' as an indicator of the amount of folding, goes from 0 to 1 with nat = 1 being the fully folded state and n = 0 being fully unfolded state. Though, in reality nat = 0 (fully extended state) is rarely physically reached by protein like polymers and the unfolded ensemble has a distribution of free energies with the mean of the unfolded well not occurring at nat=0. The real unfolded states of proteins being random heterpolymers interacting with the solvents, typically have the free energy well of the unfolded states centered on nat = 0.1-0.25 values depending on their size. In the unfolded regime, the entropic terms dwarf any favorable energetic contributions that are typically negligible. The maxima of the entropy functional (when using a parameterized value of ΔS_{res} = 17.5 J $mol^{-1}K^{-1}$, it occurs at nativeness values of nat = 0.13 for a 100 residue protein, Figure 5.5) therefore could be taken as the representative entropy of realistic unfolded states.

The conformational entropy cost of protein folding to be evaluated here is the ΔS_{tot} cost of going from the unfolded state to native state, i.e. the $S_{unf} - S_{native}$ difference between the entropies in these states. The bigger question is how to evaluate the entropy of the unfolded state ensemble that entails the oft-posed controversial subject on what exactly constitutes the unfolded state. Using the observations about the unfolded states from the free energy surface model, this controversy is circumvented.

The final total entropy cost of folding is calculated using the equation:

$$\Delta S_{max} = max\ (\ (-R[n\ln(n)+(1-n)\ln(1-n)]+n\Delta S_{res}^{n=1}+(1-n)\Delta S_{res}^{n=0})\ )\quad (5.18)$$

taking native state as the reference state and the maximum value (ΔS_{max}) of entropy in the entropy functional vs nativeness to be the total conformational cost ΔS_{conf} for the protein.

5.6.7 Estimation of Conformational Entropy for individual proteins

For any given protein with its 3D structure available, the dihedral angles are first calculated and the native clusters are assigned for each amino acid. The backbone entropic cost (ΔS_{bb}) is calculated by summing the cost of fixing each amino acid in the corresponding native cluster (using Table 5.1). Side chain entropic costs (ΔS_{sc}) as calculated using the above procedure is then added to the backbone entropic cost (ΔS_{bb}) to estimate the total conformational entropic cost (ΔS_{tot}) of folding for a given protein. Total or individual costs are normalized by the size (number of residues) of corresponding protein to calculate the per residue entropy costs. Eq. 5.18 is then used to obtain the total conformational entropic cost of folding the protein based on the free energy surface model with native state as the reference.

5.7 Results and Discussion

5.7.1 Benchmarking the theoretical results by comparison with experimental data

Using the above mentioned carefully curated dataset of proteins with experimentally characterized thermodynamic parameters based on the compilation by Robertson and Murphy[147], the structure based theoretical approach to calculate the total conformational entropic costs of protein folding is calibrated and benchmarked.

Experimental entropies given in Table 5.2 were extrapolated to 385K ($\Delta S_{exp,258}$) from the experimentally measured values using ΔH_m and ΔC_p for comparing with the predicted values, based on the equations 5.4 & 5.5

Correlation between the backbone entropies evaluated based on the φ,ψ clustering and the experimental numbers are given in Figure 5.6. The predicted backbone entropies, being major determinants of the total conformational entropic costs show a very high correlation with R-value of 0.98. The slope of the fitted line is 0.83. It shows that the φ/ψ based method for evaluating backbone entropy captures the essential signal linking structure and energetics quite well.

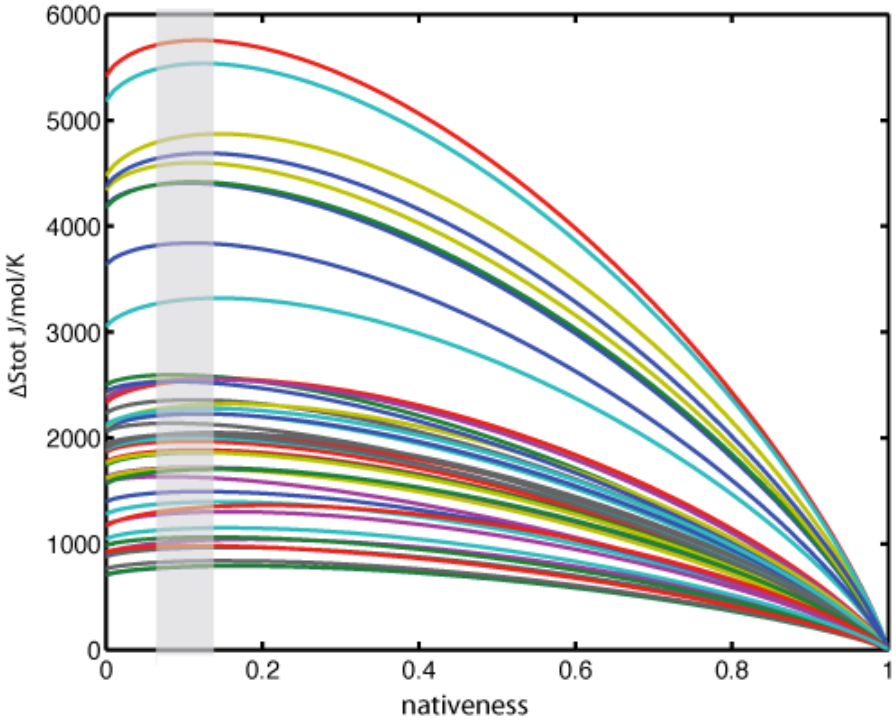


Figure 5.4 Maxima of the total entropy costs from the Free energy surface (FES) model. Position of entropy maxima on the nativeness varies between proteins and the range is shown as grey shaded area**.**

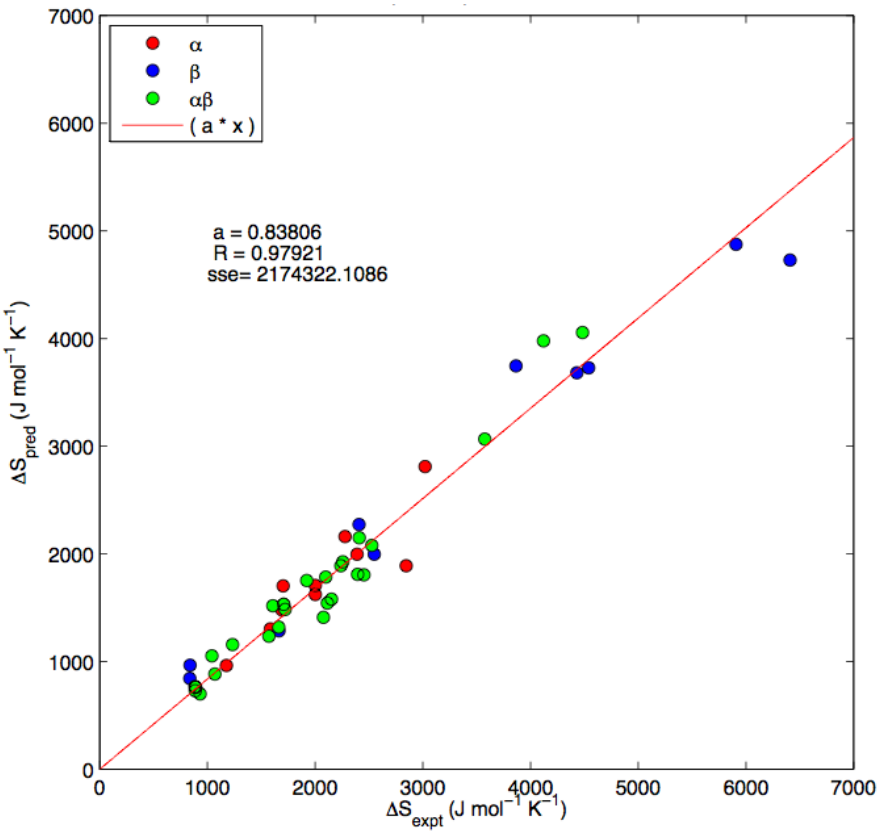


Figure 5.5 Correlation of the predicted backone entropy costs predicted based on our method and the experimentally determined entropy costs extrapolated to 385 K.

5.7.2 Per-Residue Entropies

It has been well established that many properties such as stabilities, folding rates, heat capacities etc. of proteins have a strong size scaling effect. The heteropolymeric nature of proteins determines many of their properties not only in the unfolded states but also those of the folded states. Thus it is critical to test how much of the above correlation exists beyond the simple size scaling effects. Dividing the predicted and experimental entropies for the test proteins by their corresponding number of residues (size normalization) factors out the size effects. This leads to a more stringent comparison of whether the observed correlations contain real meaning (signal) in them. In our case, the backbone entropy predictions when normalized by size yields a significant correlation of R = 0.63. More importantly there is a clear trend in the per residue entropies based on the fitted line with an intercept of 13 J $mol^{-1}K^{-1}$ and a slope of 0.25 (Figure 5.7). Another important aspect is the difference in spread of the per residue entropy costs between the experiments and the prediction. Experimental per residue numbers has a large spread with a range of 13-23.1 J $mol^{-1}K^{-1}$ and a standard deviation of 2.45 whereas the predicted numbers have a narrower range with a standard deviation of 1.2. One of the major contributors to the spread in normalized experimental numbers is the error in the experiments. For example even a small error in the measurement of the protein concentration could lead to a significant error in the thermodynamic parameters determined from those experiments. It has been estimated that 2% is the reproducibility in determining extinction coefficients of proteins used to measure the concentrations [151]. Another source of error is the propagation of the experimental values to 385 K using ΔH_m and ΔC_p values in the Robertson and Murphy dataset. For ΔH_m the reported errors range from 2%-10% and for ΔC_p the estimated errors are from 4%-10%. As each of these parameters calculated from experimental data also have some errors in them, the propagation includes their contribution as well.

ΔS_{U-F} (385) estimated using a totally different approach based on protein folding kinetics data by Akmal and Muñoz for 6 different proteins show a larger spread in the per residue entropy costs (18 ± 4 J $mol^{-1}K^{-1}$) that corroborates the fact that the experimental entropy costs tend have larger variance.

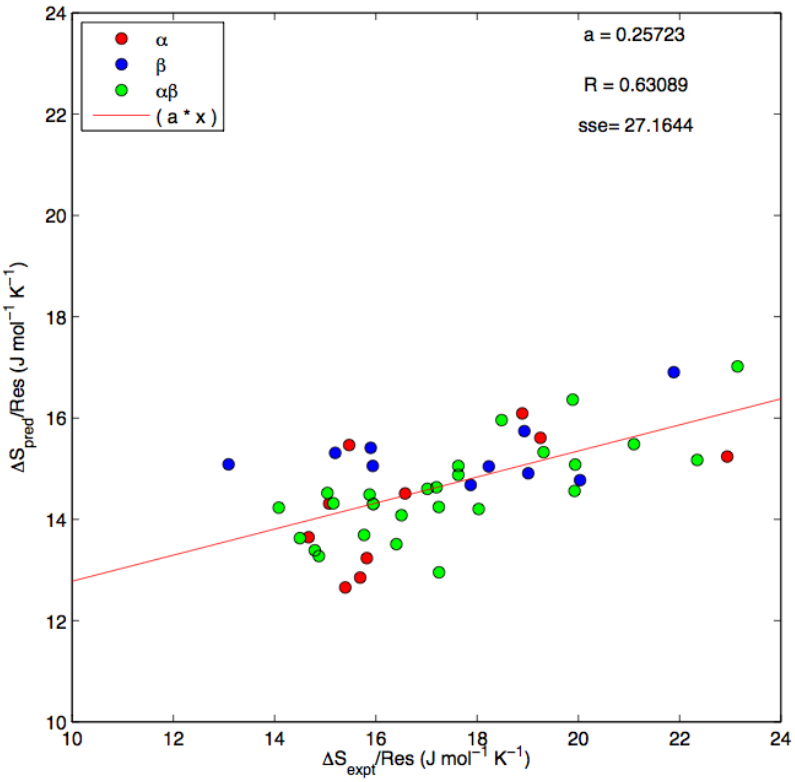


Figure 5.6 Correlation plots for size normalized entropies. Per residue entropy from prediction is plotted against those from the experiments.

5.7.3 Adding the side-chains entropic contributions

Besides the backbone contributions to the total conformational entropic costs of folding, a significant fraction of the cost is amino acid specific contributions derived from fixing their side chains. We introduce the side chain contributions for the proteins in the dataset, as explained in the methods section.

For the total entropy costs with the side chain entropies included, the correlation slightly increases to 0.67 but importantly the slope is almost 1 (a=0.99). Interestingly, the side chain contributions mainly add to the size normalized correlations of the predicted and experimental per residue entropic costs, increasing the intercept to 13.9 J $mol^{-1}K^{-1}$ with a slope of 0.47.

Table 5.2 Predicted and experimental entropic costs of proteins in $Jmol^{-1}K^{-1}$

	Name	PDB	Size	$\Delta S_{exp}(385)$	ΔS_{bb}	ΔS_{sc}	ΔS_{pred}
1	α-chymotrypsin	5cha	234	4430.29	3716.03	598.08	4314.11
2	α-chymotrypsinogen	2cga	243	3864.14	3772.87	585.21	4358.08
3	α-lactalbumin	1hml	121	1921.12	1883.73	219.57	2103.3
4	α-lactalbumin	1alc	120	2392.74	1964.89	359.56	2324.45
5	Acyl carrier protein	1t8k	75	1177.00	912.21	70.28	982.49
6	Arabinose binding protein	1abe	303	4483.03	3951.41	658.34	4609.75
7	Arc Repressor	1arr	104	2002.00	1852.9	400.71	2253.61
8	B1 domain of protein G	1pgb	54	885.95	684.77	65.71	750.48
9	B2 domain of protein G	1pgx	54	931.58	656.5	117.41	773.91
10	Barnase	1bni	106	2112.00	1551.62	342.01	1893.63
11	Barstar	1ay7	87	1568.99	1201.27	171.24	1372.51
12	BPTI	5pti	56	883.00	768.08	210.89	978.97
13	Carbonic anhydrase B	2cab	254	4539.61	3730.98	802.52	4533.5
14	Chymotrypsin inhibitor 2	1coa	62	1069.24	851.04	69.49	920.53
15	Cytochrome b5	1cyo	86	1661.08	1376.16	138.4	1514.56
16	Cytochrome c (horse)	1hrc	102	1691.03	1509.76	193.29	1703.05
17	Cytochrome c (yeast iso-1)	2pcc	106	2002.36	1904.53	222.28	2126.8
18	Cytochrome c (yeast iso-2)	1yea	110	1702.20	1939.68	202.3	2141.98
19	His containing protein	2hpr	85	1232.68	1166.47	81.19	1247.66
20	Interleukin 1-β	9ilb	151	2406.77	2229.1	233.32	2462.42
21	Lysozyme (human)	1lz1	128	2256.67	2006.51	279.82	2286.33
22	Lysozyme (hen)	1lys	127	2239.00	2010.48	269.74	2280.22
23	Lysozyme(equine)	2eql	127	2526.10	2343.95	346.88	2690.84
24	Lysozyme (T4)	2lzm	162	2410.00	2244.55	390.12	2634.67
25	Met repressor	1cmb	206	3022.02	2804.16	414.94	3219.1
26	Myoglobin(horse)	1ymb	151	2277.26	2538.87	296.19	2835.06
27	Myoglobin(whale)	1mbo	151	2389.13	1925.18	301.71	2226.89
28	Ovomucoid 3rd domain	2ovo	54	891.28	737.00	46.53	783.53
29	Papain	9pap	210	3574.13	3059.16	624.13	3683.29
30	Parvalbumin	5cpv	107	1706.51	1514.21	317.57	1831.78
31	Pepsin	5pep	324	5907.28	4781.27	835.76	5617.03
32	Pepsinogen	3psg	320	6410.56	4736.82	825.8	5562.62
33	Plaminogen K4 domain	1pmk	76	1663.39	1379	335.24	1714.24
34	RNase T1	5rnt	102	2152.06	1660.43	192.93	1853.36
35	RNase A	3rn3	122	2098.52	1789.26	267.44	2056.7
36	ROP	1rpr	124	2844.80	2088.92	470.47	2559.39
37	Sac 7d	1wd0	64	837.91	947.72	105.52	1053.24
38	α-Spectrin	1shg	55	836.00	831.24	142.41	973.65
39	Staphylococcus nuclease	1stn	134	2547.53	1981.8	254.14	2235.94
40	Stefin A	1nb5	93	1719.13	1471.58	175.31	1646.89
41	Stefin B	1stf	93	2078.31	1392.24	284.82	1677.06
42	Subtilisininhibitor	3sic	106	2453.26	1865.25	190.03	2055.28
43	Subtilisin BPN	2st1	274	4121.27	4018.94	680.92	4699.86
44	Thioredoxin	2trx	106	1607.03	1552.86	208.42	1761.28
45	Trp repressor	2wrp	103	1585.89	1176.03	36.8	1212.83
46	Ubiquitin	1ubq	74	1042.17	1036.37	107.42	1143.79

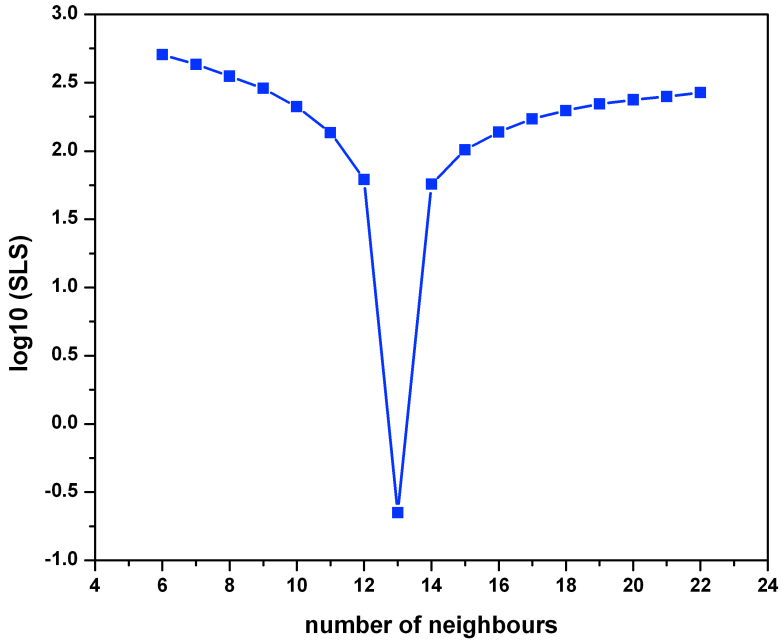


Figure 5.7 Choosing the number of neighbors cutoff for defining the restricted side chains based on minimum in the sum of least squares (SLS) comparison between predicted and experimental entropies including side chains based on different number of neighbors within 4.5Å distance criteria.

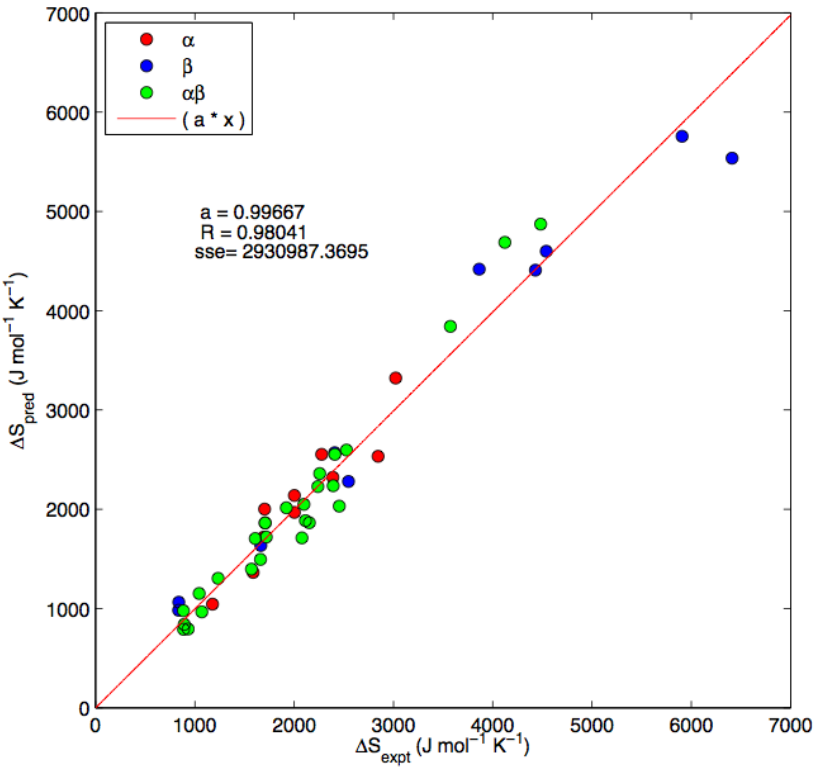


Figure 5.8 Correlation of the total conformational entropic costs including the side chain contributions. Correlation improves with the inclusion of side chain entropies.

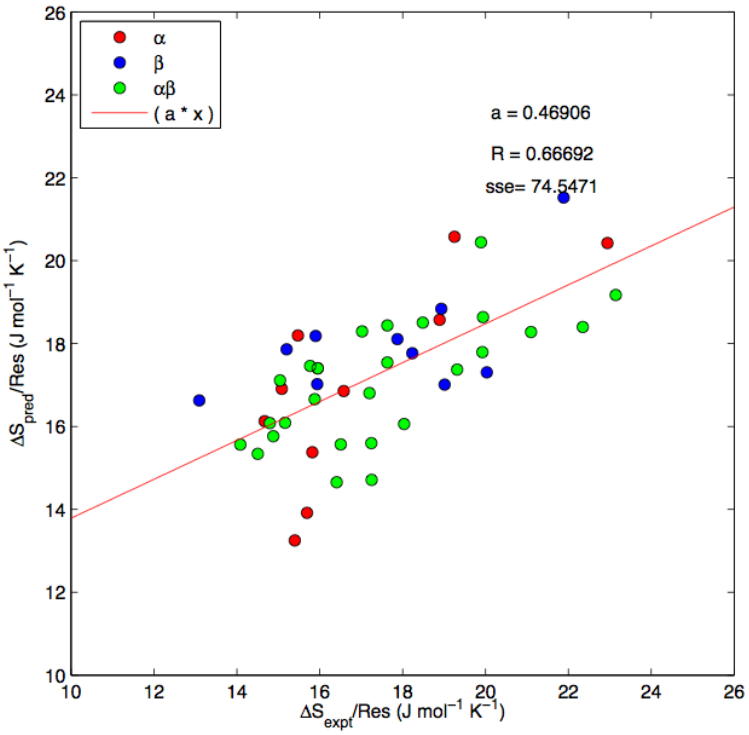


Figure 5.9 Correlation plots of size normalized entropies. Entropy from prediction is plotted against that from the experiments.

5.7.4 Predicting entropy costs for Kinetics dataset

The motivation for the development of this method is to introduce sequence and structure specificity to the mean field free energy surface model that only has broad features such as size of the protein as input. Towards that, we predict the entropy costs of a set of 52 proteins with experimentally characterized folding and unfolding kinetics and calculate their size normalized entropy costs that are shown in Table 5.3. This curated dataset of kinetics data was used to develop a method[76] for predicting absolute folding and unfolding rates based only on size and structural class of the protein as inputs using the version free energy surface model described in Chapter 4. Prediction accuracies of the method were high both for folding and unfolding rates with accuracies of ±0.7 and ±1.4 orders of magnitude relative to the experimental rates span of 6 and 8 orders of magnitude respectively. Such accuracies using the model just with size and structure type could be bettered with introducing protein specific inputs based on the sequence and structure to reach the differences caused by single point mutations (±0.34 in folding rates and ±0.7 in unfolding rates).

Table 5.3 Predicted Per residue entropy costs for kinetics dataset of proteins

	Protein	PDB	Struct.	N	ΔS_{res} J/mol/K
1	WW prototype	1e0m	β	37	18.44
2	FBP28 (W30A)	1e0l	β	37	23.47
3	Yap	1k9q	β	40	20.32
4	BBL (H166W)	2bth	α	45	16.67
5	E3BD (F166W)	1w4e	α	45	17.73
6	POB (YWLA)	1w4j	α	51	16.74
7	hTRF1	1ba5	α	53	18.11
8	cMyb	1idy	α	54	15.17
9	Engrailed HD	1enh	α	54	16.31
10	Src SH3	1rlq	β	56	19.87
11	Protein G	1pgb	α + β	56	14.66
12	a-Spectrin SH3	1shg	β	57	17.86
13	ABP1 SH3	1jo8	β	58	18.46
14	Fyn SH3	1shf	β	59	16.88
15	hRAP1	1fex	α	59	19.87
16	BDPA	1ss1	α	60	19.74
17	Protein L	2ptl	α + β	62	19.40
18	Sso7d (Y34W)	1bf4	α + β	63	16.70
19	CI2	2ci2	α + β	64	17.09
20	Sho1 SH3	2vkn	β	66	17.99
21	CspB Tm	1g6p	β	66	16.37
22	CspB Bc	1c9o	β	66	22.18
23	CspB Bs	1csp	β	67	15.89
24	EC298	1ryk	α	69	16.85
25	CspA	1mjc	β	69	19.37
26	Tendamistat	3ait	β	74	19.02
27	Ubiquitin	1ubq	α + β	76	15.56
28	RafRBD	1rfa	α + β	78	20.13
29	λ-repressor 6-85	1lmb	α	80	14.03
30	ADA2h	1o6x	α + β	81	19.16
31	Hpr	1poh	α + β	85	18.22
32	bACBP	2abd	α	86	15.48
33	PI3K SH3	1pnj	β	86	22.41
34	Im9	1imq	α	86	14.97
35	Im7	1ayi	α	86	20.90
36	Tenascin	1ten	β	90	17.04
37	9 Fibronectin III	1fnf	β	90	16.78
38	PTL9 C	1div	α + β	92	17.35
39	Twitchin	1wit	β	93	23.42
40	10 Fibronectin III	1fnf	β	94	15.68
41	U1A	1urn	α + β	96	16.42
42	L23	1n88	α + β	96	21.09
43	S6	1ris	α + β	97	15.30

44	ctAcp	2acy	α + β	98	19.82
45	mAcP	1aps	α + β	98	16.80
46	Urm1	2qjl	α + β	99	16.79
47	Src SH2	1spr	α + β	103	18.81
48	Cyt b562	1yza	α	106	19.54
49	FKBP12	1fkb	α + β	107	17.19
50	Tm1023	1j5u	α + β	125	16.70
51	Azurin (apo)	1e65	α + β	128	17.37
52	CheW	1k0s	α + β	151	24.64

5.8 Conclusions

A method to calculate the total conformational entropic cost of folding that is both structure and sequence specific has been developed and evaluated. Other structure-based approaches have been developed earlier but none of these earlier attempts have an extensive benchmarking with the experimental data. The two-step clustering of φ/ψ maps with calibration of the resulting entropy numbers with curated experimental data provides an unbiased way to evaluate the critical thermodynamic parameter – folding entropic costs.

Incorporating protein specific entropic parameters along with protein specific energy parameters into the simple free energy surface models is the next step that will increase the power of such simple FES models even more.

Conclusions

Protein folding has been studied extensively both experimentally and with computer simulations. With the recent convergence of the timescales accessible to all-atom molecular dynamics simulations and the experimental characterization of many fast-folding proteins, a new era of mutual reinforcement and iterative refinement of computational and experimental methods has begun. The advances in computing and data analysis methodologies have enabled obtaining equilibrium dynamics of fast folding proteins from single long trajectories or multiple short trajectories having many folding-unfolding transitions. Now various experimental data like equilibrium NMR chemical shift measurements, fast kinetics and single molecule data both from single molecule force spectroscopy and smFRET are being utilized routinely to test the results from extensive computer simulations and to refine the chemical force fields used to perform such simulations. In principle, the atomic simulation trajectories have all the information about the folding process at the highest resolution possible. With increasing reliability and accuracy of the force fields MD simulations will eventually serve as *computational experiments* offering unprecedented and detailed access to the protein folding mechanisms that are not typically accessible to most other experiments.

On the other hand, many well-grounded simple theoretical models have been developed for protein folding offering insights about the process. They have been essential to understand many important facets of the problem from addressing how proteins fold fast (and not in astronomical timescales) to the effects of mutations on stability and kinetics of proteins. Such theoretical models typically involve making many simplifying assumptions and in fact rely on them. However, testing such simplistic assumptions behind these successful models for their validity or falsifying them experimentally has been difficult so far. There have been very few attempts to compare and test them with properties measured actually in experiments. With the advances in atomistic simulations now there is also scope for directly examining the simple models with the folding mechanisms observed in them.

Stochastic kinetic simulations act as the required bridge between the theory, simple models and experiments in protein folding including the results from full atom computational experiments. In this thesis, we have established how stochastic kinetic simulations based on simple models of protein folding offer mechanistic insights and corroborate many experimental findings.

New experimental results in helix-coil kinetics, in particular complexities of the local dynamics in helices that give rise to both simple and stretched exponential kinetics were compared with single molecule trajectories of stochastic kinetic simulations with a simple model based on nucleation-elongation. These unbiased simulations confirm the findings from the experiments and reconcile the experimental results with well-established theory of nucleation-elongation.

Local diffusive dynamics occurring in the 50-60 ns timescales in alpha helices that were measured in the experiments could be directly observed in the stochastic kinetic simulations. From the individual molecular trajectories, different rapid diffusive motions happening on the helices could be identified and such motions were labeled as the 'waltzing' of the helices

Applications of stochastic simulations to unravel the dynamics of single molecule behavior of protein were demonstrated next. From simple chemical kinetic models of two-state folding to the Brownian dynamics in a harmonic well of one state proteins, stochastic kinetic simulations helped probe their properties and establish the effects of energetic barriers on protein folding. How even smaller barriers such as ~1kT profoundly alter the dynamics and behaviors of proteins is illustrated in the stochastic trajectories of single molecule fluctuations.

With the identification and experimental characterization of an increasing number of proteins with zero or low energetic barriers in the past years, the necessity to understand their implications towards specific biological functions also increases. Using stochastic simulations in combination with simple models offer a quick way to achieving this.

Fast folding proteins that have low or zero barriers or activated kinetics push the boundaries of many experimental techniques and equally the quantitative methods of analysis of data produced from such experiments. For smFRET experiments that are on the forefront of single molecule studies of protein folding, we have developed a rigorous procedure for directly extracting the free energy surfaces and conformational dynamics from the time stamped photon arrival sequences measured in these experiments. By combining the simple free energy surface models with a maximum likelihood method, our new procedure performs photon by photon analysis of the experimental data to identify the right folding scenario and important parameters of folding including the barrier heights and the dynamic intramolecular diffusion coefficient D. Again, we demonstrate the application and usefulness of stochastic kinetic simulations in the development of the procedure. Using stochastic simulations we synthetically generate trajectories of single molecule fluctuations on the 1D free energy surfaces and model photon arrival sequences to test the procedure extensively. We calibrate the performance of the method based on quantity and quality of the available photon arrival data including the number of available photons, ratio of interphoton arrival times and the amount of noise in the data.
We demonstrate that the procedure performs reliably well even with small number of photons, with low ratios of photon arrival times to the relaxation time of the protein and also with an amount of background noise levels in the data.

Relating protein structure with energetics by developing a method to evaluate protein folding entropic costs from the 3D structure is another important result from this thesis. The method is based on a novel clustering of the φ/ψ space and evaluating entropic costs of fixing particular amino acids to the clusters using a simple formula. Side chain entropies are calculated based on determining the restriction of residues depending on the number of contacting neighbors and calculating the costs of fixing the side chain in the particular rotameric state.

Simple free energy surface model is then used to locate the entropy maxima to evaluate the folding entropy costs of the protein. The method is calibrated with a curated set of experimental measurements and the performance is benchmarked. High correlations are observed between the predicted and experimental values. The method is put to a further stringent test by making comparisons after factoring out the size scaling effects and established to perform well. With this method, the sequence specific amino acid effects could be introduced into the simple free energy surface model and the scope of the model tremendously expanded.

Summary of Results

Establishing simple stochastic kinetic simulations as a way to unravel and demystify behaviors of dynamic molecules.

Developing a procedure using combined simple models of protein folding and maximum likelihood analysis for directly obtaining the free energy surfaces and conformational dynamics from photon by photon analysis of smFRET data.

Statistical analysis of protein structures to establish relationships between protein structure and energetics and incorporating structure specific entropic costs in the simple free energy surface model of proteins.

Conclusiones

El plegamiento de las proteínas ha sido ampliamente estudiado, tanto experimentalmente como mediante simulaciones computacionales. Con la reciente convergencia de, por un lado, el acceso a escalas de tiempo mayores con las simulaciones de dinámica molecular a escala atómica y, por otro lado, la caracterización experimental de numerosas proteínas de plegamiento rápido, ha dado comienzo una nueva era de fortalecimiento mutuo y de perfeccionamiento iterativo entre los métodos computacionales y experimentales. Los avances en técnicas computacionales y en metodologías de análisis de datos han permitido obtener la dinámica en equilibrio de proteínas con plegamiento rápido a partir de trayectorias individuales largas o trayectorias múltiples cortas con gran cantidad de transiciones plegamiento-desplegamiento. Hoy en día, diferentes datos experimentales, entre los que se incluyen medidas del desplazamiento químico en equilibrio por RMN, cinéticas rápidas y experimentos de molécula única obtenidos a partir de espectroscopia de fuerza atómica o de FRET, están siendo utilizados rutinariamente para corroborar los resultados adquiridos a partir de extensas simulaciones computacionales y para refinar los campos de fuerza químicos utilizados en la realización de dichas simulaciones. En principio, las trayectorias generadas en las simulaciones atómicas poseen toda la información sobre el proceso de plegamiento a la mayor resolución posible. Un incremento de la fiabilidad y la precisión de los campos de fuerza de las simulaciones de dinámica molecular permitirá que estas simulaciones eventualmente actúen como *experimentos computacionales* que ofrecerán un acceso a detalles de los mecanismos del plegamiento de proteínas no accesibles para la mayor parte de las técnicas experimentales.

Por otra parte, se han desarrollado modelos teóricos simples de plegamiento de proteínas sólidamente fundamentados que han arrojado luz sobre los mecanismos del proceso. Dichos modelos han sido esenciales para entender importantes aspectos del proceso, desde abordar el tema de cómo las proteínas se pliegan rápidamente (y no en escalas astronómicas de tiempo) hasta los efectos de las mutaciones en la estabilidad y la cinética de las proteínas. Estos modelos teóricos se realizan normalmente basándose en muchas asunciones que simplifican los propios modelos y, de hecho, dependen de ellos. Sin embargo, la confirmación experimental de estas asunciones simplificadoras que están detrás de los modelos exitosos ha sido difícil hasta ahora. De hecho, se han realizado escasos intentos de comparar y confirmar los modelos mediante el uso de parámetros obtenidos experimentalmente. Gracias a los avances en las simulaciones a nivel atómico, existe actualmente interés en examinar directamente los modelos simples con los mecanismos de plegamiento que implican.

Las simulaciones de cinética estocástica actúan como un puente necesario entre la teoría, los modelos simples y los experimentos en el campo del plegamiento de proteínas, incluyendo los resultados obtenidos a partir de experimentos computacionales a nivel atómico. En esta tesis, hemos establecido cómo las

simulaciones de cinética estocástica basadas en modelos simples de plegamiento ofrecen información mecanística y corroboran muchos descubrimientos experimentales.

Hemos comparado nuevos resultados cinéticos experimentales de transiciones tipo hélice-ovillo, en particular experimentos relacionados con la complejidad de la dinámica local en hélices que dan origen tanto a cinéticas exponenciales simples como estiradas, con trayectorias de molécula única de simulaciones de cinética estocástica mediante un modelo simple tipo nucleación-elongación. Estas simulaciones imparciales confirman los hallazgos experimentales y concilian los resultados experimentales con la teoría establecida de nucleación-elongación.

La dinámica de difusión local existente en el rango de tiempo de 50-60 ns en alfa hélices obtenida experimentalmente pudo observarse directamente mediante las simulaciones de cinética estocástica. A partir de las trayectorias de moléculas individuales se identificaron diferentes movimientos rápidos en las hélices, que fueron denominados como el 'Vals' de las hélices.

Tras esto, hemos demostrado la aplicación de las simulaciones estocásticas para desvelar el comportamiento dinámico de las proteínas a nivel de molécula única. Desde modelos químicos cinéticos simples para procesos de plegamiento de dos estados hasta la dinámica browniana en un pozo harmónico para proteínas de un estado, las simulaciones de cinética estocástica han ayudado a confirmar sus propiedades y establecer los efectos de las barreras energéticas en el plegamiento de proteínas. A partir de las trayectorias estocásticas de las fluctuaciones de una única molécula se han conseguido ofrecer datos incluso de cómo las menores barreras energéticas –aproximadamente de 1kT– modifican profundamente la dinámica y el comportamiento de las proteínas.

La identificación y caracterización experimental de un número cada vez mayor de proteínas con nulas o bajas barreras energéticas en los últimos años ha incrementado la necesidad de entender sus implicaciones en relación a sus funciones biológicas específicas. La aplicación de simulaciones estocásticas en combinación con modelos simples ofrece una rápida solución para lograr el objetivo anterior.

Las proteínas con plegamiento rápido que poseen bajas o nulas barreras o cinéticas activadas sobrepasan los límites de muchas técnicas experimentales y, del mismo modo, los de los métodos cuantitativos de análisis de datos obtenidos de esos experimentos. En el caso de los experimentos smFRET, en primera línea dentro de los estudios de molécula única, hemos desarrollado un rigoroso procedimiento para extraer directamente las superficies de energía libre y la dinámica conformacional a partir de los tiempos de llegada de las secuencias de fotones medidas en estos experimentos. Combinando modelos simples de superficies de energía libre con un método de máxima probabilidad, nuestro nuevo procedimiento realiza un análisis de los datos experimentales fotón a fotón con el fin de identificar el escenario de plegamiento correcto, así como importantes parámetros del plegamiento, incluyendo las alturas de las barreras

energéticas y el coeficiente *D* de difusión dinámica intramolecular. De nuevo, demostramos la aplicación de las simulaciones de cinética estocástica en el desarrollo del procedimiento. Utilizando simulaciones estocásticas hemos generado sintéticamente tanto trayectorias de fluctuaciones de una única molécula dentro de la superficies de energía libre de una dimensión y como secuencias modelo de llegada de fotones para comprobar ampliamente el procedimiento. Hemos calibrado el funcionamiento del método en base a la cantidad y calidad de los datos de llegada de fotones disponibles, variando el número de fotones disponibles, el ratio de tiempo de llegada entre fotones y la cantidad de ruido en los datos.

Asimismo, hemos demostrado que el procedimiento posee una gran fiabilidad incluso con un pequeño número de fotones, con bajos ratios entre el tiempo de llegada de fotones y el tiempo de relajación de la proteína, y con diferentes niveles de ruido en los datos.

Otro importante resultado de esta tesis es el desarrollo de un método que relaciona la estructura de las proteínas con la energética mediante la evaluación del coste entrópico del plegamiento a partir de la estructura tridimensional de la proteína. El método se basa en un novedoso agrupamiento del espacio Phi/Psi y en la evaluación de los costes entrópicos de la fijación angular de determinados aminoácidos para aparecer en cada uno de las agrupaciones utilizando una fórmula simple. La entropía de las cadenas laterales se calcula basándose en la determinación de la restricción de los residuos dependiendo del número de residuos vecinos con los que están en contacto y en el cálculo de los costes de fijar la cadena lateral en un estado rotacional en particular. Un modelo simple de superficies de energía libre es entonces utilizado para localizar el máximo de entropía para evaluar los costes entrópicos del plegamiento de la proteína. El método está calibrado con un conjunto de medidas experimentales cuidadosamente seleccionado y el funcionamiento ha sido evaluado comparativamente. La comparación entre los valores experimentales y las predicciones demuestra una gran correlación entre ambos. El método ha sido sometido exitosamente a otro riguroso test donde se han realizado comparativas tras una eliminación de los efectos del escalado por tamaño. Con este método se pudieron introducir en el modelo simple de superficies de energía libre los efectos de la secuencia específica de aminoácidos y, así, expandir tremendamente el alcance del modelo.

Resumen de los resultados

Establecimiento de las simulaciones de cinética estocástica simples como un método para desentrañar y desmitificar comportamientos de moléculas dinámicas.

Desarrollo de un procedimiento, combinando modelos simples de plegamiento de proteínas y análisis de máximos de probabilidad, para obtener directamente las superficies de energía libre y la dinámica conformacional a partir de análisis fotón a fotón de los datos obtenidos mediante smFRET.

Análisis estadístico de las estructuras proteicas para establecer relaciones entre la estructura y la energética de las proteínas e incorporación de los costes entrópicos relacionados con la estructura en los modelos simples de superficies de energía libre de proteín

List of Publications

List of Publications from and during the Thesis work

1. Waltzing α-helices. Victor Munoz and Ravishankar Ramanathan, Proceedings of the National Academy of Sciences Vol. 106, 1299-1300 (2009)

2. Are there Still Surprises Buried Inside Statistical Analysis of Protein Structure? Verma A, Ravishankar Ramanathan, Journal of Biomol. Structure and Dynamics Vol. 28, 661-662 (2011)

3. A photoprotection strategy for microsecond-resolution single-molecule fluorescence spectroscopy. Luis Campos, Liu J, Wang X, Ravishankar Ramanathan, English D & Victor Munoz. Nature Methods 8, 143-146 (2011)

4. Exploring one-state downhill protein folding in single molecules. Liu J, Luis Campos, Cerminara M, Wang X, Ravishankar Ramanathan, English D & Victor Munoz. Proceedings of the National Academy of Sciences Vol. 109, 179-184 (2012)

5. Slow proton transfer coupled to unfolding explains the puzzling results of single molecule experiments on BBL, a paradigmatic downhill folding protein. Cerminara M, Luis Campos, Ravishankar Ramanathan & Victor Muñoz. PLOS One 8, e78044 (2013)

Manuscripts Under Preparation:

1. Decoding conformational dynamics and free energy surfaces of fast-folding proteins from single molecule Photon Arrival Trajectories Ravishankar Ramanathan and Victor Muñoz (Submitted to Journal of Physical Chemistry B)

2. Statistics of amino acid distributions in the φ/ψ Dihedral angles Space and derivation of Entropic Costs of Protein Folding Ravishankar Ramanathan, Abhinav Verma and Victor Muñoz (To be submitted to Proteins: Structure Function and Bioinformatics)

Bibliography

1. Schrodinger, E., What's Life? The Physical Aspect ofthe Living Cell. Cambridge University Press, Cambridge: 1945.
2. Mulder, G. J., On the composition of some animal substances. *Journal für praktische Chemie* **1839,** *16* (129), 15.
3. Watson, J. D.; Crick, F. H., Molecular structure of nucleic acids. *Nature* **1953,** *171* (4356), 737-738.
4. Anfinsen, C. B., Principles that govern folding of protein chains. *Science* **1973,** *181* (4096), 223-230.
5. Ebbinghaus, S.; Dhar, A.; McDonald, J. D.; Gruebele, M., Protein folding stability and dynamics imaged in a living cell. *Nature methods* **2010,** *7* (4), 319-323.
6. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A., The sequence of the human genome. *science* **2001,** *291* (5507), 1304-1351.
7. Koboldt, D. C.; Steinberg, K. M.; Larson, D. E.; Wilson, R. K.; Mardis, E. R., The next-generation sequencing revolution and its impact on genomics. *Cell* **2013,** *155* (1), 27-38.
8. Gilbert, J. A.; Dupont, C. L., Microbial metagenomics: beyond the genome. *Annual Review of Marine Science* **2011,** *3*, 347-371.
9. Ecker, J. R.; Bickmore, W. A.; Barroso, I.; Pritchard, J. K.; Gilad, Y.; Segal, E., Genomics: ENCODE explained. *Nature* **2012,** *489* (7414), 52-55.
10. Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S., Towards a knowledge-based human protein atlas. *Nature biotechnology* **2010,** *28* (12), 1248-1250.
11. Uhlen, M. The Human Protein Atlas. http://www.proteinatlas.org (accessed November 2014).
12. Church, G. M.; Elowitz, M. B.; Smolke, C. D.; Voigt, C. A.; Weiss, R., Realizing the potential of synthetic biology. *Nature Reviews Molecular Cell Biology* **2014**.
13. Tompa, P.; Fersht, A., *Structure and function of intrinsically disordered proteins*. CRC Press: 2010.
14. Editorial: So much more to know. Science , 309:78-102. *Science* **2005,** *309*, 78-102.
15. Dill, K. A.; MacCallum, J. L., The Protein-Folding Problem, 50 Years On. *Science* **2012,** *338* (6110), 1042-1046.
16. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G., Funnels, pathways, and the energy landscape of protein folding - a synthesis. *Proteins-Structure Function and Genetics* **1995,** *21* (3), 167-195.
17. Muñoz, V., *Protein folding, misfolding and aggregation: Classical themes and novel approaches*. Royal Society of Chemistry: **2008**; Vol. 13.
18. Socci, N. D.; Onuchic, J. N.; Wolynes, P. G., Diffusive dynamics of the reaction coordinate for protein folding funnels. *Journal of Chemical Physics* **1996,** *104* (15), 5860-5868. 19. Onuchic, J. N.; LutheySchulten, Z.; Wolynes, P. G., Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry* **1997,** *48*, 545-600.

20. Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Munoz, V., Experimental identification of downhill protein folding. *Science* **2002,** *298* (5601), 2191-2195.
21. Sadqi, M.; Fushman, D.; Munoz, V., Atom-by-atom analysis of global downhill protein folding. *Nature* **2006,** *442* (7100), 317-321.
22. Liu, J.; Campos, L. A.; Cerminara, M.; Wang, X.; Ramanathan, R.; English, D. S.; Munoz, V., Exploring one-state downhill protein folding in single molecules. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109* (1), 179-184.
23. (a) Fung, A.; Li, P.; Godoy-Ruiz, R.; Sanchez-Ruiz, J. M.; Munoz, V., Expanding the realm of ultrafast protein folding: gpW, a midsize natural single-domain with alpha+beta topology that folds downhill. *Journal of the American Chemical Society* **2008,** *130* (23), 7489-7495; (b) Sborgi, L.; Verma, A.; Munoz, V.; de Alba, E., Revisiting the NMR Structure of the Ultrafast Downhill Folding Protein gpW from Bacteriophage lambda. *Plos One* **2011,** *6* (11).
24. Sanchez-Medina, C.; Sekhar, A.; Vallurupalli, P.; Cerminara, M.; Munoz, V.; Kay, L. E., Probing the Free Energy Landscape of the Fast-Folding gpW Protein by Relaxation Dispersion NMR. *Journal of the American Chemical Society* **2014,** *136* (20), 7444-7451.
25. Gelman, H.; Gruebele, M., Fast protein folding kinetics *Quarterly Reviews of Biophysics* **2014,** *47* (02), 1469-8994.
26. Munoz, V., Conformational dynamics and ensembles in protein folding. In *Annual Review of Biophysics and Biomolecular Structure*, 2007; Vol. 36, pp 395-412.
27. Lapidus, L. J.; Steinbach, P. J.; Eaton, W. A.; Szabo, A.; Hofrichter, J., Effects of chain stiffness on the dynamics of loop formation in polypeptides. Appendix: Testing a 1-dimensional diffusion model for peptide dynamics. *The Journal of Physical Chemistry B* **2002,** *106* (44), 11628-11640.
28. Kubelka, J.; Hofrichter, J.; Eaton, W. A., The protein folding 'speed limit'. *Current Opinion in Structural Biology* **2004,** *14* (1), 76-88.
29. Yang, W. Y.; Gruebele, M., Rate-temperature relationships in lambda-repressor fragment lambda 6-85 folding. *Biochemistry* **2004,** *43* (41), 13018-25.
30. Yang, W. Y.; Gruebele, M., Folding at the speed limit. *Nature* **2003,** *423* (6936), 193-197.
31. Liu, F.; Gruebele, M., Tuning lambda(6-85) towards downhill folding at its melting temperature. *Journal of Molecular Biology* **2007,** *370* (3), 574-584.
32. Borgia, A.; Williams, P. M.; Clarke, J., Single-Molecule Studies of Protein Folding. *Annual Review of Biochemistry* **2008,** *77* (1), 101-125.
33. (a) Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A., Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science* **2012,** *335* (6071), 981-984; (b) Chung, H. S.; Eaton, W. A., Single-molecule fluorescence probes dynamics of barrier crossing. *Nature* **2013,** *502* (7473), 685-+.
34. Cecconi, C.; Shank, E. A.; Bustamante, C.; Marqusee, S., Direct observation of the three-state folding of a single protein molecule. *Science* **2005,** *309* (5743), 2057-2060.
35. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How Fast-Folding Proteins Fold. *Science* **2011,** *334* (6055), 517-520.

36. Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S., To milliseconds and beyond: challenges in the simulation of protein folding. *Current opinion in structural biology* **2013,** *23* (1), 58-65.
37. (a) Go, N., Theoretical studies of protein folding. *Annual review of biophysics and bioengineering* **1983,** *12* (1), 183-210; (b) Onuchic, J. N.; Wolynes, P. G., Theory of protein folding. *Current Opinion in Structural Biology* **2004,** *14* (1), 70-75.
38. Orozco, M.; Orellana, L.; Hospital, A.; Naganathan, A. N.; Emperador, A.; Carrillo, O.; Gelpí, J., Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. *Adv Protein Chem Struct Biol* **2011,** *85*, 183-215.
39. Dill, K. A.; Bromberg, S.; Yue, K.; Chan, H. S.; Ftebig, K. M.; Yee, D. P.; Thomas, P. D., Principles of protein folding—a perspective from simple exact models. *Protein Science* **1995,** *4* (4), 561-602.
40. Mirny, L.; Shakhnovich, E., Protein folding theory: from lattice to all-atom models. *Annual review of biophysics and biomolecular structure* **2001,** *30* (1), 361-396.
41. Bryngelson, J. D.; Wolynes, P. G., Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry* **1989,** *93* (19), 6902-6915.
42. Plaxco, K. W.; Simons, K. T.; Baker, D., Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of Molecular Biology* **1998,** *277* (4), 985-994.
43. (a) Zwanzig, R., Simple model of protein folding kinetics. *Proceedings of the National Academy of Sciences* **1995,** *92* (21), 9801-9804; (b) Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A., A statistical mechanical model for beta-hairpin kinetics. *Proceedings of the National Academy of Sciences of the United States of America* **1998,** *95* (11), 5872-5879.
44. Munoz, V.; Eaton, W. A., A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (20), 11311-11316.
45. Bruscolini, P.; Pelizzola, A., Exact solution of the Munoz-Eaton model for protein folding. *Physical review letters* **2002,** *88* (25), 258101.
46. (a) Naganathan, A. N., Predictions from an Ising-like Statistical Mechanical Model on the Dynamic and Thermodynamic Effects of Protein Surface Electrostatics. *Journal of Chemical Theory and Computation* **2012,** *8* (11), 4646-4656; (b) Naganathan, A. N., A Rapid, Ensemble and Free Energy Based Method for Engineering Protein Stabilities. *Journal of Physical Chemistry B* **2013,** *117* (17), 4956-4964.
47. (a) Alm, E.; Baker, D., Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (20), 11305-11310; (b) Alm, E.; Morozov, A. V.; Kortemme, T.; Baker, D., Simple physical models connect theory and experiment in protein folding kinetics. *Journal of molecular biology* **2002,** *322* (2), 463-476.
48. Galzitskaya, O. V.; Finkelstein, A. V., A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (20), 11299-11304.

49. Munoz, V., Thermodynamics and kinetics of downhill protein folding investigated with a simple statistical mechanical model. *International Journal of Quantum Chemistry* **2002,** *90* (4-5), 1522-1528.
50. Naganathan, A. N.; Doshi, U.; Fung, A.; Sadqi, M.; Munoz, V., Dynamics, energetics, and structure in protein folding. *Biochemistry* **2006,** *45* (28), 8466-8475.
51. Naganathan, A. N.; Doshi, U.; Munoz, V., Protein folding kinetics: Barrier effects in chemical and thermal denaturation experiments. *Journal of the American Chemical Society* **2007,** *129* (17), 5673-5682.
52. Levitt, M.; Warshel, A., Computer simulation of protein folding. *Nature* **1975,** *253* (5494), 694-698.
53. Best, R. B., Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology* **2012,** *22* (1), 52-61.
54. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C., Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM* **2008,** *51* (7), 91-97.
55. Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W., Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010,** *330* (6002), 341-346.
56. Gillespie, D. T., Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* **1977,** *81* (25), 2340-2361.
57. Gillespie, D. T., Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **2007,** *58*, 35-55.
58. Muñoz, V.; Serrano, L., Helix design, prediction and stability. *Current opinion in biotechnology* **1995,** *6* (4), 382-386.
59. Munoz, V.; Serrano, L., Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* **1997,** *41* (5), 495-509.
60. Schwarz, G., On the kinetics of the helix-coil transition of polypeptides in solution. *Journal of molecular biology* **1965,** *11* (1), 64-77.
61. Gruenewald, B.; Nicola, C.; Lustig, A.; Schwarz, G.; Klump, H., Kinetics of the helix—coil transition of a polypeptide with non-ionic side groups, derived from ultrasonic relaxation measurements. *Biophysical chemistry* **1979,** *9* (2), 137-147.
62. Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B., Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry* **1996,** *35* (3), 691-697.
63. Thompson, P. A.; Eaton, W. A.; Hofrichter, J., Laser temperature jump study of the helix⇌ coil kinetics of an alanine peptide interpreted with a'kinetic zipper'model. *Biochemistry* **1997,** *36* (30), 9200-9210.
64. Thompson, P. A.; Munoz, V.; Jas, G. S.; Henry, E. R.; Eaton, W. A.; Hofrichter, J., The helix-coil kinetics of a heteropeptide. *Journal of Physical Chemistry B* **2000,** *104* (2), 378-389.
65. Huang, C.-Y.; Getahun, Z.; Zhu, Y.; Klemke, J. W.; DeGrado, W. F.; Gai, F., Helix formation via conformation diffusion search. *Proceedings of the National Academy of Sciences* **2002,** *99* (5), 2788-2793.

66. Doshi, U. R.; Munoz, V., The principles of alpha-helix formation: Explaining complex kinetics with nucleation-elongation theory. *Journal of Physical Chemistry B* **2004,** *108* (24), 8497-8506.
67. (a) Hummer, G.; García, A. E.; Garde, S., Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins: Structure, Function, and Bioinformatics* **2001,** *42* (1), 77-84; (b) Sorin, E. J.; Pande, V. S., Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical Journal* **2005,** *88* (4), 2472-2493.
68. Lapidus, L. J.; Eaton, W. A.; Hofrichter, J., Measuring dynamic flexibility of the coil state of a helix-forming peptide. *Journal of molecular biology* **2002,** *319* (1), 19-25.
69. Fierz, B.; Reiner, A.; Kiefhaber, T., Local conformational dynamics in α-helices measured by fast triplet transfer. *Proceedings of the National Academy of Sciences* **2009,** *106* (4), 1057-1062.
70. Doshi, U.; Munoz, V., Kinetics of alpha-helix formation as diffusion on a one-dimensional free energy surface. *Chemical Physics* **2004,** *307* (2-3), 129-136.
71. Muñoz, V.; Ramanathan, R., Waltzing α-helices. *Proceedings of the National Academy of Sciences* **2009,** *106* (5), 1299-1300.
72. Jackson, S. E.; Fersht, A. R., Folding of chymotrypsin inhibitor 2. 1. Evidence for a 2-state transition. *Biochemistry* **1991,** *30* (43), 10428-10435.
73. Callen, H. B.; Welton, T. A., Irreversibility and generalized noise. *Physical Review* **1951,** *83* (1), 34.
74. Campos, L. A.; Liu, J.; Wang, X.; Ramanathan, R.; English, D. S.; Munoz, V., A photoprotection strategy for microsecond-resolution single-molecule fluorescence spectroscopy. *Nature Methods* **2011,** *8* (2), 143-U63.
75. Henry, E. R.; Best, R. B.; Eaton, W. A., Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **2013,** *110* (44), 17880-17885.
76. De Sancho, D.; Muñoz, V., Integrated prediction of protein folding and unfolding rates from only size and structural class. *Physical Chemistry Chemical Physics* **2011,** *13* (38), 17030-17043.
77. Naganathan, A. N.; Perez-Jimenez, R.; Sanchez-Ruiz, J. M.; Munoz, V., Robustness of downhill folding: Guidelines for the analysis of equilibrium folding experiments on small proteins. *Biochemistry* **2005,** *44* (20), 7435-7449.
78. Naganathan, A. N.; Munoz, V., Scaling of folding times with protein size. *Journal of the American Chemical Society* **2005,** *127* (2), 480-481.
79. Naganathan, A. N.; Munoz, V., Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** *107* (19), 8611-8616.
80. Li, P.; Oliva, F. Y.; Naganathan, A. N.; Munoz, V., Dynamics of one-state downhill protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (1), 103-108.
81. Naganathan, A. N.; Munoz, V., Thermodynamics of Downhill Folding: Multi-Probe Analysis of PDD, a Protein that Folds Over a Marginal Free Energy Barrier. *Journal of Physical Chemistry B* **2014,** *118* (30), 8982-8994.
82. Naganathan, A. N.; Li, P.; Perez-Jimenez, R.; Sanchez-Ruiz, J. M.; Munoz, V., Navigating the Downhill Protein Folding Regime via Structural Homologues. *Journal of the American Chemical Society* **2010,** *132* (32), 11183-11190.

83. Schuler, B., Single-molecule FRET of protein structure and dynamics - a primer. *Journal of Nanobiotechnology* **2013,** *11* (Suppl 1), S2.
84. (a) Lakowicz, J. R., *Principles of Fluorescence Spectroscopy*. 1999; (b) Lipman, E. A.; Schuler, B.; Bakajin, O.; Eaton, W. A., Single-molecule measurement of protein folding kinetics. *Science* **2003,** *301* (5637), 1233-1235.
85. Stryer, L.; Haugland, R. P., Energy transfer - a spectroscopic ruler. *Proceedings of the National Academy of Sciences of the United States of America* **1967,** *58* (2), 719-&.
86. Ha, T.; Enderle, T.; Ogletree, D. F.; Chemla, D. S.; Selvin, P. R.; Weiss, S., Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences of the United States of America* **1996,** *93* (13), 6264-6268.
87. Ratner, V.; Kahana, E.; Eichler, M.; Haas, E., A general strategy for site-specific double labeling of globular proteins for kinetic FRET studies. *Bioconjugate Chemistry* **2002,** *13* (5), 1163-1170.
88. Panchuk-Voloshina, N.; Haugland, R. P.; Bishop-Stewart, J.; Bhalgat, M. K.; Millard, P. J.; Mao, F.; Leung, W. Y., Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *Journal of Histochemistry & Cytochemistry* **1999,** *47* (9), 1179-1188.
89. Rhoades, E.; Gussakovsky, E.; Haran, G., Watching proteins fold one molecule at a time. *Proceedings of the National Academy of Sciences of the United States of America* **2003,** *100* (6), 3197-3202.
90. Eggeling, C.; Berger, S.; Brand, L.; Fries, J.; Schaffer, J.; Volkmer, A.; Seidel, C., Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. *Journal of biotechnology* **2001,** *86* (3), 163-180.
91. Gopich, I. V.; Szabo, A., Theory of Single Molecule FRET Efficiency Histograms. *Single-Molecule Biophysics: Experiment and Theory, Volume 146* **2011**, 245-297.
92. Sisamakis, E.; Valeri, A.; Kalinin, S.; Rothwell, P. J.; Seidel, C. A. M., Accurate single-molecule fret studies using multiparameter fluorescence detection. In *Methods in Enzymology, Vol 475: Single Molecule Tools, Pt B: Super-Resolution, Particle Tracking, Multiparameter, and Force Based Methods*, Walter, N. G., Ed. 2010; Vol. 475, pp 455-514.
93. Gopich, I. V.; Szabo, A., Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109* (20), 7747-7752.
94. (a) Selvin, P. R.; Ha, T., *Single-molecule techniques: a laboratory manual*. Cold Spring Harbour Laboratory Press 2008; (b) Hinterdorfer, P.; van Oijen, A. M., *Handbook of single molecule biophysics*. Springer: 2009.
95. Schuler, B.; Hofmann, H., Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales. *Current Opinion in Structural Biology* **2013,** *23* (1), 36-47.
96. Schuler, B.; Lipman, E. A.; Eaton, W. A., Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **2002,** *419* (6908), 743-747.

97. Nettels, D.; Mueller-Spaeth, S.; Kuester, F.; Hofmann, H.; Haenni, D.; Rueegger, S.; Reymond, L.; Hoffmann, A.; Kubelka, J.; Heinz, B.; Gast, K.; Best, R. B.; Schuler, B., Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (49), 20740-20745.
98. Brucale, M.; Schuler, B.; Samorì, B., Single-Molecule Studies of Intrinsically Disordered Proteins. *Chemical reviews* **2014**.
99. Blanco, M.; Walter, N. G., Analysis of complex single-molecule fret time trajectories. In *Methods in Enzymology, Vol 472: Single Molecule Tools, Pt A: Fluorescence Based Approaches*, Walter, N. G., Ed. 2010; Vol. 472, pp 153-178.
100. Antonik, M.; Felekyan, S.; Gaiduk, A.; Seidel, C. A. M., Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis. *Journal of Physical Chemistry B* **2006,** *110* (13), 6970-6978.
101. Kalinin, S.; Valeri, A.; Antonik, M.; Felekyan, S.; Seidel, C. A. M., Detection of Structural Dynamics by FRET: A Photon Distribution and Fluorescence Lifetime Analysis of Systems with Multiple States. *Journal of Physical Chemistry B* **2010,** *114* (23), 7983-7995.
102. Nir, E.; Michalet, X.; Hamadani, K. M.; Laurence, T. A.; Neuhauser, D.; Kovchegov, Y.; Weiss, S., Shot-noise limited single-molecule FRET histograms: comparison between theory and experiments. *The Journal of Physical Chemistry B* **2006,** *110* (44), 22103-22124.
103. Gopich, I. V.; Szabo, A., Single-macromolecule fluorescence resonance energy transfer and free-energy profiles. *The Journal of Physical Chemistry B* **2003,** *107* (21), 5058-5063.
104. Gopich, I.; Szabo, A., Theory of photon statistics in single-molecule Forster resonance energy transfer. *Journal of Chemical Physics* **2005,** *122* (1).
105. Gopich, I. V.; Szabo, A., Single-molecule FRET with diffusion and conformational dynamics. *Journal of Physical Chemistry B* **2007,** *111* (44), 12925-12932.
106. Chung, H. S.; Gopich, I. V., Fast single-molecule FRET spectroscopy: theory and experiment. *Physical chemistry chemical physics : PCCP* **2014,** *16* (35), 18644-57.
107. Aitken, C. E.; Marshall, R. A.; Puglisi, J. D., An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments. *Biophys J* **2008,** *94* (5), 1826-35.
108. Michalet, X.; Colyer, R. A.; Scalia, G.; Ingargiola, A.; Lin, R.; Millaud, J. E.; Weiss, S.; Siegmund, O. H. W.; Tremsin, A. S.; Vallerga, J. V.; Cheng, A.; Levi, M.; Aharoni, D.; Arisaka, K.; Villa, F.; Guerrieri, F.; Panzeri, F.; Rech, I.; Gulinatti, A.; Zappa, F.; Ghioni, M.; Cova, S., Development of new photon-counting detectors for single-molecule fluorescence microscopy. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2013,** *368* (1611).
109. Schroder, G. F.; Grubmuller, H., Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments. *Journal of Chemical Physics* **2003,** *119* (18), 9920-9924.
110. Milescu, L. S.; Yildiz, A.; Selvin, P. R.; Sachs, F., Maximum likelihood estimation of molecular motor kinetics from staircase dwell-time sequences. *Biophysical Journal* **2006,** *91* (4), 1156-1168.

111. Milescu, L. S.; Akk, G.; Sachs, F., Maximum likelihood estimation of ion channel kinetics from macroscopic currents. *Biophysical Journal* **2005,** *88* (4), 2494-2515.
112. Chung, H. S.; Gopich, I. V.; McHale, K.; Cellmer, T.; Louis, J. M.; Eaton, W. A., Extracting Rate Coefficients from Single-Molecule Photon Trajectories and FRET Efficiency Histograms for a Fast-Folding Protein. *Journal of Physical Chemistry A* **2011,** *115* (16), 3642-3656.
113. Haas, K. R.; Yang, H.; Chu, J. W., Expectation-Maximization of the Potential of Mean Force and Diffusion Coefficient in Langevin Dynamics from Single Molecule FRET Data Photon by Photon. *The journal of physical chemistry. B* **2013,** *117* (49), 15591-605.
114. Andrec, M.; Levy, R. M.; Talaga, D. S., Direct determination of kinetic rates from single-molecule photon arrival trajectories using hidden Markov models. *The Journal of Physical Chemistry A* **2003,** *107* (38), 7454-7464.
115. Lee, T.-H., Extracting Kinetics Information from Single-Molecule Fluorescence Resonance Energy Transfer Data Using Hidden Markov Models. *Journal of Physical Chemistry B* **2009,** *113* (33), 11535-11542.
116. Keller, B. G.; Kobitski, A.; Jaschke, A.; Nienhaus, G. U.; Noe, F., Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *J Am Chem Soc* **2014,** *136* (12), 4534-43.
117. Liu, Y.; Park, J.; Dahmen, K. A.; Chemla, Y. R.; Ha, T., A Comparative Study of Multivariate and Univariate Hidden Markov Modelings in Time-Binned Single-Molecule FRET Data Analysis. *The Journal of Physical Chemistry B* **2010,** *114* (16), 5386-5403.
118. Jung, S.; Dickson, R. M., Hidden Markov Analysis of Short Single Molecule Intensity Trajectories. *The Journal of Physical Chemistry B* **2009,** *113* (42), 13886-13890.
119. Kou, S. C.; Sunney Xie, X.; Liu, J. S., Bayesian analysis of single-molecule experimental data. *Journal of the Royal Statistical Society Series C-Applied Statistics* **2005,** *54*, 469-496.
120. Gopich, I. V.; Szabo, A., Decoding the Pattern of Photon Colors in Single-Molecule FRET. *Journal of Physical Chemistry B* **2009,** *113* (31), 10965-10973.
121. Chung, H. S.; Cellmer, T.; Louis, J. M.; Eaton, W. A., Measuring ultrafast protein folding rates from photon-by-photon analysis of single molecule fluorescence trajectories. *Chemical Physics* **2013,** *422*, 229-237.
122. Chung, H. S.; Louis, J. M.; Eaton, W. A., Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (29), 11837-11844.
123. De Sancho, D.; Doshi, U.; Munoz, V., Protein folding rates and stability: how much is there beyond size? *Journal of the American Chemical Society* **2009,** *131* (6), 2074-2075.
124. Ramachandran, G.; Ramakrishnan, C. t.; Sasisekharan, V., Stereochemistry of polypeptide chain configurations. *Journal of molecular biology* **1963,** *7* (1), 95-99.
125. Pauling, L.; Corey, R. B.; Branson, H. R., The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* **1951,** *37* (4), 205-211.

126. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M., PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography* **1993,** *26* (2), 283-291.
127. Stein, E. G.; Rice, L. M.; Brünger, A. T., Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *Journal of Magnetic Resonance* **1997,** *124* (1), 154-164.
128. Murphy, K. P.; Murphy, K. P.; Murphy, K. P., *Protein structure, stability, and folding*. Springer: 2001; Vol. 117.
129. Schellman, J. A., The stability of hydrogen-bonded peptide structures in aqueous solution. *Comptes rendus des travaux du Laboratoire Carlsberg. Série chimique* **1955,** *29* (14-15), 230.
130. Némethy, G.; Scheraga, H. A., Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method. *Biopolymers* **1965,** *3* (2), 155-184.
131. Wang, J.; Purisima, E. O., Analysis of thermodynamic determinants in helix propensities of nonpolar amino acids through a novel free energy calculation. *Journal of the American Chemical Society* **1996,** *118* (5), 995-1001.
132. D'Aquino, J. A.; Gómez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E., The magnitude of the backbone conformational entropy change in protein folding. *Proteins: Structure, Function, and Bioinformatics* **1996,** *25* (2), 143-156.
133. Doig, A. J.; Sternberg, M. J., Side‐chain conformational entropy in protein folding. *Protein Science* **1995,** *4* (11), 2247-2251.
134. Lee, K. H.; Xie, D.; Freire, E.; Amzel, L. M., Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. *Proteins: Structure, Function, and Bioinformatics* **1994,** *20* (1), 68-84.
135. Privalov, P. L., Thermodynamics of protein folding. *The Journal of Chemical Thermodynamics* **1997,** *29* (4), 447-474.
136. Zhang, C.; Cornette, J. L.; Delisi, C., Consistency in structural energetics of protein folding and peptide recognition. *Protein science* **1997,** *6* (5), 1057-1064.
137. Yang, D.; Mok, Y.-K.; Forman-Kay, J. D.; Farrow, N. A.; Kay, L. E., Contributions to protein entropy and heat capacity from bond vector motions measured by NMR spin relaxation. *Journal of molecular biology* **1997,** *272* (5), 790-804.
138. Alexandrescu, A. T.; Rathgeb‐Szabo, K.; Jahnke, W.; Schulthess, T.; Kammerer, R. A.; Rumpel, K., 15N backbone dynamics of the S‐peptide from ribonuclease A in its free and S‐protein bound forms: Toward a site‐specific analysis of entropy changes upon folding. *Protein science* **1998,** *7* (2), 389-402.
139. Fitter, J., A measure of conformational entropy change during thermal protein unfolding using neutron spectroscopy. *Biophysical journal* **2003,** *84* (6), 3924-3930.
140. Thompson, J. B.; Hansma, H. G.; Hansma, P. K.; Plaxco, K. W., The backbone conformational entropy of protein folding: experimental measures from atomic force microscopy. *Journal of Molecular biology* **2002,** *322* (3), 645-652.
141. Kauzmann, W., Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry* **1959,** *14*, 1-63.
142. (a) Tanford, C., Protein denaturation. *Advances in protein chemistry* **1968,** *23*, 121-282; (b) Baldwin, R. L., Temperature dependence of the hydrophobic

interaction in protein folding. *Proceedings of the National Academy of Sciences* **1986,** *83* (21), 8069-8072.
143. Privalov, P.; Khechinashvili, N., A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *Journal of molecular biology* **1974,** *86* (3), 665-684.
144. Murphy, K.; Gill, S., Group additivity thermodynamics for dissolution of solid cyclic dipeptides into water. *Thermochimica acta* **1990,** *172*, 11-20.
145. Murphy, K. P.; Freire, E., Thermodynamics of structural stability and cooperative folding behavior in proteins. *Advances in protein chemistry* **1992,** *43*, 313-361.
146. Akmal, A.; Munoz, V., The nature of the free energy barriers to two-state folding. *Proteins-Structure Function and Bioinformatics* **2004,** *57* (1), 142-152.
147. Robertson, A. D.; Murphy, K. P., Protein structure and the energetics of protein stability. *Chemical Reviews* **1997,** *97* (5), 1251-1267.
148. Wang, G.; Dunbrack, R. L., PISCES: recent improvements to a PDB sequence culling server. *Nucleic acids research* **2005,** *33* (suppl 2), W94-W98.
149. Boomsma, W.; Mardia, K. V.; Taylor, C. C.; Ferkinghoff-Borg, J.; Krogh, A.; Hamelryck, T., A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences* **2008,** *105* (26), 8932-8937.
150. Dunbrack Jr, R. L.; Karplus, M., Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of molecular biology* **1993,** *230* (2), 543-574.
151. Pace, C. N.; Vajdos, F.; Fee, L.; Grimsley, G.; Gray, T., How to measure and predict the molar absorption coefficient of a protein. *Protein science* **1995,** *4* (11), 2411-2423.